Applications du calcul des probabilités à la recherche de régions génomiques conservées

Simona Grusea

UNIVERSITÉ DE PROVENCE
U.F.R. M.I.M.
ÉCOLE DOCTORALE DE MATHÉMATIQUES ET INFORMATIQUE E.D. 184

THÈSE

présentée pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE PROVENCE

Spécialité : Mathématiques

par

Simona GRUSEA

sous la direction de Pr. Etienne PARDOUX et Dr. Pierre PONTAROTTI

Titre :

Applications du calcul des probabilités à la recherche de régions genomiques conservées

soutenue publiquement le 3 décembre 2008

JURY

M. Yuri GOLUBEV	Université de Provence	Examinateur
M. Ralph NEININGER	J. W. Goethe Universitat	Rapporteur
M. Etienne PARDOUX	Université de Provence	Directeur de thèse
M. Pierre PONTAROTTI	Université de Provence	Codirecteur de thèse
Mme. Sophie SCHBATH	INRA, Jouy en Josas	Rapporteur
M. Anton WAKOLBINGER	J. W. Goethe Universitat	Examinateur

Remerciements

C'est avec beaucoup de joie que je souhaite remercier mon directeur de thèse Etienne Pardoux, pour tout ce qu'il m'a apporté pendant ces années. Pour sa grande disponibilité, pour la confiance et la liberté qu'il m'a accordées, pour son soutien inconditionné, pour l'esprit de perfection qu'il m'a appris, pour sa grande intuition qui m'a fait sortir beaucoup de fois de l'impasse, pour sa gentillesse (même quand il me grondait), pour l'attention avec laquelle il a corrigé tous mes manuscrits, pour ses petites leçons de grammaire et pour son visage toujours lumineux, comme le ciel de Marseille.

Je voudrais également remercier mon co-directeur de thèse Pierre Pontarotti, pour tout son soutien et sa disponibilité pendant ces années, pour sa grande passion pour la recherche et la persévérance qu'il m'a apprise et pour m'avoir fait découvrir, petit à petit, un peu des mystères de la vie.

Je tiens à remercier mes deux rapporteurs Sophie Schbath et Ralph Neininger pour l'attention qu'ils ont accordée à mon manuscrit et pour toutes leurs questions et remarques.

Je remercie également tous les membres du jury d'avoir accepté de participer à ma soutenance.

Je souhaite dire un grand merci ("mulţumesc") à tous mes collègues et amis, pour leur soutien et leur amitié, pour partager avec moi, pendant toutes ces années, des joies et des difficultés, et pour me laisser de très beaux souvenirs.

Je dédie cette thèse à mon mari Pompiliu, à mes parents et à ma soeur Adriana, qui ont été loin des yeux parfois, mais toujours près du coeur. Je leur remercie pour tout.

Table des matières

Introduction

Le point de départ de cette thèse est le stage de master recherche que j'ai effectué, sous la direction d'Etienne Pardoux et la co-direction de Pierre Pontarotti, au Laboratoire Evolution Biologique et Modélisation de l'Université de Provence. C'est ici que j'ai commencé à decouvrir les joies de la recherche et les mystères de la science de la vie, dont je voudrais vous faire part.

Cette thèse se concentre autour de quelques sujets de probabilités et statistique liés à la génomique comparative. Le sujet majeur est la détection de régions génomiques conservées entre espèces.

L'identification de régions génomiques qui sont conservées entre différentes espèces est un premier pas essentiel dans l'essai de déchiffrer l'histoire évolutive des espèces et sert aussi pour mieux comprendre la biologie des espèces modernes.

Les génomes des êtres vivants sont eux aussi "vivants" : ils évoluent continûment, en subissant des changements dans leur contenu en gènes (par duplications, pertes de gènes, transferts horizontaux) et aussi des changements dans l'ordre des gènes, suite à des réarrangements à grande échelle, comme les inversions, les translocations, les transpositions, les fusions et les fissions chromosomiques.

Les gènes qui descendent d'un même gène ancestral suite à un évènement de spéciation s'appellent *gènes orthologues*. En général ils gardent la même fonction. En conséquence, les clusters de gènes orthologues peuvent être un signe pour leur provenance d'une même région ancestrale et pour la proximité évolutive des deux espèces, mais aussi pour l'existence d'une pression de sélection fonctionnelle agissant sur les gènes dans le cluster.

On a besoin de tests statistiques pour distinguer les clusters orthologues significatifs des groupements des gènes qui ont pu apparaître dans le génome par hasard, au cours du temps, suite aux différents types d'évènements génomiques.

Dans la littérature il y a plusieurs définitions des clusters de gènes (voir [6, 11, 14, 21, 22, 23, 29]). Une des plus connues est la notion de "max-gap cluster" ([22, 23, 29]), où l'on restreint la longuer des écarts entre les orthologues consécutifs dans le cluster à un seuil fixé. Une autre notion est celle d'*intervalle commun* ("common interval", voir [21]), qui est un ensemble de gènes orthologues se trouvant consécutivement, mais pas nécéssairement dans le même ordre, dans deux ou plusieurs espèces. Une généralisation de la notion d'intervalle commun est celle de "gene team" (voir [6]), dans laquelle la restriction de consécutivité est relâchée, à condition que les écarts entre les orthologues ne dépassent pas un certain seuil.

Pour pouvoir détecter des signaux évolutifs même entre des espèces très éloignées, nous choisissons pour les régions génomiques conservées une définition très peu restrictive, comme dans [11] : *deux régions chromosomiques, dans deux espèces différentes, ayant en commun un certain nombre de gènes orthologues (aucune restriction a priori sur les écarts*

entre les orthologues ou sur leur ordre).

Une région génomique conservée est *significative* (i.e. "vraiment conservée") s'il est très improbable qu'elle soit apparue par hasard dans un génome aléatoire.

Pour simplifier, on va voir un génome comme une séquence ordonnée de gènes, sans séparation en chromosomes et on va mesurer la longueur d'une région génomique en nombre de gènes.

Il y a trois approches différentes pour chercher des régions génomiques conservées entre deux espèces, suivant la dimension de l'espace de recherche (voir [14]). Un type de recherche est de balayer les deux génomes en entier ("whole genome comparison" en anglais). Une deuxième approche est l'approche de type "région de référence", qui consiste à partir d'une région génomique fixée dans une espèce (appelée *région de référence*) et à balayer en entier le génome d'une deuxième espèce pour trouver des régions orthologues à la région de référence. Le troisième type d'approche consiste à comparer seulement deux fenêtres génomique fixées dans les deux espèces pour chercher des groupes de gènes orthologues en commun ("window-sampling approach" en anglais).

Dans chacun des trois cas la dimension de l'espace de recherche est différente et les tests statistiques appliqués pour tester la significativité des clusters trouvés doivent être adaptés au type d'approche correspondant.

Dans notre travail nous considérons le cas des régions génomiques conservées trouvées par une approche de type région de référence, dans laquelle on part d'une région fixée dans une certaine espèce A et on balaie le génome d'une autre espèce B pour trouver des régions similaires (orthologues) à la région de référence.

L'hypothèse nulle que l'on considère dans ce travail est l'hypothèse :

$$H_0 : \text{ordre aléatoire des gènes dans le génome B.}$$

Toutes les probabilités et les lois de probabilités que nous considérons dans la suite sont sous-entendues sous l'hypothèse H_0.

Test statistique basé sur la proximité des orthologues

Dans le Chapitre 1 nous présentons une approximation de Poisson composée pour calculer des probabilités impliquées dans des tests statistiques pour la significativité des régions génomiques conservées. Dans ce chapitre nous prenons en compte seulement la proximité des gènes orthologues dans les clusters.

Un aspect important de notre démarche est le fait de prendre en compte l'existence des familles multigéniques, i.e. le fait que pour un gène donné de la région de référence on peut trouver plusieurs orthologues dans le génome B (appelés *co-orthologues*), à cause d'évènements de duplication qui ont pu survenir après la séparation des deux espèces.

L'existence des familles multigéniques est un facteur important qui doit être pris en compte quand on teste la significativité des clusters de gènes orthologues, mais très peu des tests statistiques existant le considèrent. Danchin et Pontarotti [11] proposent de pondérer les orthologues en proportion inverse des tailles des familles multigéniques dont ils font partie, mais leur utilisation d'une loi binomiale n'est pas adéquate dans ce cadre. Raghupathy et Durand [29] prennent aussi en compte l'existence des familles multigéniques, mais leur test statistique est adapté seulement au cas des clusters trouvés par une approche de type "window-sampling" et non pas à notre approche de type région de référence.

Dans ce travail nous adoptons l'idée de Danchin et Pontarotti [11] pour prendre en compte les familles multigéniques.

Le Chapitre 1 est structuré comme suit.

La Section 1.1 décrit le contexte biologique. Dans la Section 1.2 nous présentons le cadre mathématique et le modèle mathématique simplifié que nous utilisons. Nous donnons la formulation mathématique du problème et nous commençons, pour des raisons techniques, par considérer le cas des génomes circulaires.

Dans la Section 1.3 nous donnons une courte présentation de la méthode de Stein-Chen pour l'approximation de Poisson composée – l'approche par couplage. Le résultat principal est le Théorème 1.3, due à Roos [30], qui donne un résultat de convergence pour l'erreur de l'approximation, sous l'hypothèse de l'existence d'un certain couplage.

La Section 1.4 est le coeur du Chapitre 1, contenant l'approximation de Poisson composée pour notre probabilité, avec le résultat de convergence que nous avons obtenu en utilisant la méthode de Stein-Chen. En suivant l'approche de Roos [30, 31], nous construisons explicitement le couplage décrit dans le Théorème 1.3 et nous estimons les termes d'erreur apparaissant dans le Théorème 1.3. Le Théorème 1.10 contient notre résultat de convergence.

Une simplification technique de nos calculs est due à l'hypothèse énoncée dans "Assumption 1", dans laquelle nous supposons que les familles multigéniques dans le génome B qui sont de taille minimale constituent une partie significative de toutes les familles multigéniques de B.

Dans la sous-section 1.4.2 nous décrivons une approximation "markovienne" pour le calcul, en pratique, des paramètres de la loi de Poisson composée.

Dans la sous-section 1.4.3 nous étendons les résultats au cas des génomes linéaires.

Dans la Section 1.5 nous présentons quelques résultats numériques, tant pour le cas circulaire que pour le cas linéaire, sur des jeux de valeurs pour les paramètres qui sont intéressants dans notre cadre biologique.

Nous avons contribué aussi à l'implémentation de notre approximation de Poisson composée dans la plateforme multi-agent C.A.S.S.I.O.P.E.

Dans la suite nous donnons une courte présentation de notre approche.

Modélisation mathématique

Les données du problème sont les suivantes :

- m : le nombre de génes de la région de référence ayant des orthologues dans B ;
- ϕ_i : le nombre d'orthologues dans B du gène i de la région de référence, $i = 1, ..., m$;
- N : le nombre total de gènes dans le génome B.

Soit $n := \phi_1 + \cdots + \phi_m$ le nombre total d'orthologues, dans le génome B, des gènes de la région de référence.

Comme on est dans le cas $m \ll N$, on va simplifier encore le modèle et on va voir le génome B comme l'intervalle $[0, 1]$ et, sous l'hypothèse nulle, les positions dans B des n orthologues comme des v.a. i.i.d. uniformément distribuées dans $[0, 1]$.

Soit $\{U_{ij}, j = 1, ..., \phi_i, i = 1, ..., m\}$ des v.a. indépendantes et uniformément distribuées sur $[0, 1]$, représentant les positions dans B des orthologues appartenant à chacune des m familles multigéniques.

Pour prendre en compte l'existence des familles multigéniques dans le génome B on utilise la mesure de comptage suivante :

$$\mu_m = \sum_{i=1}^{m} \frac{1}{\phi_i} \sum_{j=1}^{\phi_i} \delta_{U_{ij}},$$

où δ_x représente la mesure de Dirac au point x.

Pour chaque intervalle $I \subset [0,1]$, on appelle $\mu_m(I)$ son *poids*. Dans le cas sans famille multigénique (i.e. $\phi_i = 1, \forall i = 1, ..., m$), $\mu_m(I)$ est simplement le nombre d'orthologues qui se trouvent dans I.

Pour un gène appartenant à une famille multigénique de taille j, on dit qu'il a l'*étiquette* $1/j$.

Soit un poids h fixé, de la forme

$$h = \sum_{i=1}^{m} \frac{n_i}{\phi_i}, \text{ avec } 0 \leq n_i \leq \phi_i, \forall i = 1, ..., m,$$

et soit une longueur r fixée aussi.

Soit $U_{(1)} \leq U_{(2)} \leq \cdots \leq U_{(n)}$ les statistiques d'ordre des $\{U_{ij}, j = 1, ..., \phi_i, i = 1, ..., m\}$.

On note W_m la variable aléatoire qui représente le nombre de clusters dans B qui sont de poids plus grand que h et de longueur plus petite que r :

$$W_m = \sum_{k=1}^{n} \mathbf{1}_{A_k},$$

où $A_k = \{\mu_m([U_{(k)}, U_{(k)} + r]) \geq h\}$ est l'évènement d'avoir un tel cluster commençant avec le k-ème orthologue, en position $U_{(k)}$.

Dans le premier chapitre nous utilisons une approximation de Poisson composée pour calculer, pour un poids h et une longueur r donnés, la probabilité, sous l'hypothèse nulle, de trouver quelque part dans la génome B une région de poids plus grand que h et de longueur plus petite que r, i.e. $\mathbb{P}(W_m \geq 1) = \mathbb{P}(\bigcup_k A_k)$.

Approximation de Poisson composée pour $\mathbb{P}(W_m \geq 1)$

On est dans le cas d'une somme d'indicatrices qui sont dans une dépendance locale ("short-range dependence, long-range independence" en anglais). A cause de la dépendance forte entre les indicatrices qui sont très proches, les évènements A_k ont la tendance d'arriver en groupes ("clumps" en anglais). En conséquence, il semble raisonable d'approcher la loi de W_m par une loi de Poisson composée, avec un "bon" choix des paramètres.

Nous avons obtenu des résultats de convergence pour l'erreur de notre approximation en utilisant la méthode de Stein-Chen pour l'approximation de Poisson composée, introduite par Barbour, Chen et Loh ([4]).

Plus précisément, nous utilisons l'approche "par couplage" développée par Roos ([30, 31]), qui permet d'obtenir des bornes pour l'erreur de l'approximation de Poisson composée, sous la condition de l'existence d'un certain couplage (voir les Théorèmes 1.1, 1.2 et 1.3).

Nous avons construit explicitement le couplage demandé par ces théorèmes et nous avons estimé les termes d'erreur qui apparaissent dans le Théorème 1.3.

Pour énoncer les résultats obtenus nous avons besoin de quelques notations. Pour la simplicité, on ne va énoncer les résultats que dans le cas d'un génome circulaire.

On va reparamétrer le problème et on va noter par $\phi'_1 < \cdots < \phi'_J$ toutes les valeurs différentes des tailles des familles multigéniques $\phi_1, ..., \phi_m$, et par $g_j = \left|\{i = 1, ..., m : \phi_i = \phi'_j\}\right|$, $j = 1, ..., J$, leurs multiplicités.

Remarque. *On peut représenter la mesure μ_m comme*

$$\mu_m = \sum_{i=1}^{n} L_i \delta_{U_{(i)}},$$

où $\boldsymbol{L} = (L_1, ..., L_n)$ est un vecteur aléatoire indépendant des U_i et uniformément distribué sur l'ensemble

$$\Lambda = \left\{ \boldsymbol{\ell} = (\ell_1, ..., \ell_n) \in \left\{ \frac{1}{\phi'_1}, ..., \frac{1}{\phi'_J} \right\}^n : \left| \left\{ i : \ell_i = \frac{1}{\phi'_j} \right\} \right| = g_j \phi'_j, \forall j \right\}$$

de toutes les étiquetages possibles pour les n orthologues.

Soit $\phi'_1 := \min\{\phi_i : i = 1, ..., m\}$ et soit $n_{\min} := |\{i : \phi_i = \phi'_1\}|$ le nombre de familles multigéniques dans B qui sont de taille minimale.

Soit $h_* := \lceil h\phi'_1 \rceil$. Nous supposons que $n_{\min} \geq h_*$, de telle sorte que h_* soit le nombre minimal d'orthologues dans un cluster de poids plus grand que h.

Pour tout étiquetage $\boldsymbol{\ell} \in \Lambda$ et pour tout $k = 1, ..., n$, soit

$$h_k(\boldsymbol{\ell}) := \min\{d : \ell_k + \cdots + \ell_{k+d-1} \geq h\}$$

le nombre minimal d'orthologues qu'un cluster commençant avec le k-ème orthologue doit contenir pour être de poids plus grand que h.

On a donc

$$A_k \cap \{\boldsymbol{L} = \boldsymbol{\ell}\} = \{U_{(k+h_k(\boldsymbol{\ell})-1)} - U_{(k)} \leq r\} \cap \{\boldsymbol{L} = \boldsymbol{\ell}\}.$$

On note aussi

$$h^* := \max_{\boldsymbol{\ell}} \{h_1(\boldsymbol{\ell})\}.$$

Exemple.

Supposons que dans la région de référence on ait $m = 10$ gènes qui ont au moins un orthologue dans l'espèce B. Supposons que parmi ces dix gènes, six ont un seul orthologue, trois ont deux orthologues et un gène a trois orthologues.

Dans cet exemple le vecteur des tailles différentes des familles multigéniques est $\boldsymbol{\phi}' = [1, 2, 3]$, le vecteur des multiplicités est $\boldsymbol{g} = [6, 3, 1]$, $n = 14, n_{min} = 6$ et Λ est l'ensemble de toutes les étiquetages différentes pour les 14 gènes orthologues dans le génome B. On a six gènes avec l'étiquette 1, six avec l'étiquette $\frac{1}{2}$ et trois avec l'étiquette $\frac{1}{3}$, donc Λ est l'ensemble de toutes les permutations avec répétition (multipermutations) dans lesquelles 1 apparaît 6 fois, $\frac{1}{2}$ apparaît 6 fois et $\frac{1}{3}$ apparaît 3 fois.

Supposons que $h = 3, 5$.

Le nombre minimal de gènes que doit contenir un cluster pour être de poids plus grand que h dépend de l'étiquetage des gènes.

Par exemple, si l'étiquetage est $\boldsymbol{\ell} = (1, 1, 1, 1, 1, 1, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, alors la taille minimale d'un cluster qui commence avec le premier gène et est de poids plus grand que h est égale à 4 (il suffit de prendre dans le cluster les quatre premiers gènes qui sont de poids 1). Dans ce cas $h_1(\boldsymbol{\ell}) = 4$.

Par contre, si l'étiquetage est $\boldsymbol{\ell} = (1, \frac{1}{2}, \frac{1}{3}, 1, 1, 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{2}, 1, \frac{1}{2}, 1, \frac{1}{3}, \frac{1}{2}, \frac{1}{2})$, alors $h_1(\boldsymbol{\ell}) = 5$.

Le nombre h_* représente la plus petite taille possible d'un cluster de poids plus grand que h, et correspond au "meilleurs des cas", quand le cluster est construit seulement avec des gènes appartenant à une famille de taille minimale ϕ_1' (i.e. des gènes de poids maximal) – dans notre cas, avec des gènes de poids 1. Dans cet exemple h_* vaut 4.

Le nombre h^* correspond au "pire des cas", quand le cluster doit contenir le maximum de gènes pour dépasser le poids h. Ici $h^* = 8$ et correspond, par exemple, à l'étiquetage $\boldsymbol{\ell} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 1, 1, 1, 1, 1, 1)$.

Pour tout $k = 1, ..., n$, on note I_k la fonction indicatrice de l'évènement A_k.

On va approcher la loi de W_m par une loi de Poisson composée avec les paramètres suivants :

$$\hat{\lambda}_i = \frac{1}{i} \sum_{k=1}^{n} \mathbb{E}(I_k \mathbf{1}_{\{Z_k = i\}}), \ i = 2, ..., h_* - 1,$$
$$\hat{\lambda}_1 = \mathbb{E}(W_m) - \sum_{i=2}^{h_*-1} i\hat{\lambda}_i,$$

où

$$Z_k = \sum_{j=k-h_*+2}^{k+h_*-2} I_j$$

représente le nombre d'évènements qui se réalisent dans le voisinage "de dépendance forte" de l'évènement A_k.

On approche donc notre probabilité d'intérêt $\mathbb{P}(W_m \geq 1)$ par la probabilité correspondante pour la loi de Poisson composée, i.e.

$$p := 1 - \exp\{- \sum_{i=1}^{h_*-1} \hat{\lambda}_i\}.$$

Nous avons obtenu le résultat de convergence suivant.

Théorème (Theorem 1.10)**.** *Supposons que $n \longrightarrow \infty, r \longrightarrow 0$ de telle façon que $nr \longrightarrow 0$ et $n_{\min} \asymp n$. Alors, uniformément en*

$$\frac{1}{n} \leq nr < 1 \text{ et } n > 2(2h_* + h^* - 4) \vee \exp\left\{\frac{4(2h_* + h^* - 4)}{3(h_* - 1) + h^*}\right\},$$

on a :

$$|\mathbb{P}(W_m \geq 1) - p| \leq C \frac{n(nr)^{2(h_*-1)}}{[(h_* - 1)!]^2}(1 + \mathcal{O}(\frac{1}{n}) + \mathcal{O}(nr)),$$

où

$$C = 4h^* - h_* - 6 + (h_* - 1)\{2h_* + h^* - 5 + 2^{h_*-2}(h_* + h^* - 4)\} + (h_* - 2)2^{2h_*-6}.$$

De plus, si $\mathbb{E}(W_m) = \pi_\infty$ *est constante quand* $n \longrightarrow \infty$*, alors*

$$|\mathbb{P}(W_m \geq 1) - p| = \mathcal{O}(\frac{1}{n}).$$

Le calcul des paramètres

Nous allons montrer que les termes dominants dans l'expression de $\hat{\lambda}_i$ sont ceux qui contiennent des produits de i indicatrices consécutives (voir la preuve du Lemme 1.9). On va donc utiliser l'approximation suivante :

$$\begin{aligned}\hat{\lambda}_i &\approx \frac{1}{i}\sum_{k=1}^{n}\{\mathbb{E}[(1-I_{k-i})I_{k-i+1}I_{k-i+2}\cdots I_{k-1}I_k(1-I_{k+1})] \\ &\quad + \cdots + \mathbb{E}[(1-I_{k-1})I_k\cdots I_{k+i-1}(1-I_{k+i})]\} \\ &= n\mathbb{E}[(1-I_1)I_2\cdots I_{i+1}(1-I_{i+2})] \\ &= n\mathbb{P}(A_1^C \cap A_2\cdots A_{i+1}\cap A_{i+2}^C).\end{aligned}$$

En utilisant ensuite une approximation "markovienne" :

$$\mathbb{P}(A_1^C \cap A_2\cdots A_{i+1}\cap A_{i+2}^C) \approx \mathbb{P}(A_1^C|A_2)\mathbb{P}(A_2)\mathbb{P}(A_3|A_2)\cdots\mathbb{P}(A_{i+1}|A_i)\mathbb{P}(A_{i+2}^C|A_{i+1}),$$

on obtient

$$\begin{aligned}\hat{\lambda}_i &\approx n\pi q^{i-1}(1-q)^2, \text{ pour } i = 2, ..., h_* - 1 \\ \hat{\lambda}_1 &= n\pi - \sum_{i=2}^{h_*-1} i\hat{\lambda}_i,\end{aligned}$$

où

$$\begin{aligned}\pi &= \mathbb{P}(A_1), \\ q &= \mathbb{P}(A_2|A_1).\end{aligned}$$

Pour calculer π et q on fait la somme sur tous les étiquetages possibles $\boldsymbol{\ell}$.

Le problème des tests multiples

Quand on cherche des régions génomiques conservées significatives, on ne fixe pas à l'avance le poids h de la région, ce qui entraîne un problème de tests multiples.

Une idée que nous avons eu pour tenir compte de ce problème est la suivante.

Pour un h donné, soit L_h la longueur de la plus petite région dans B qui est de poids plus grand que h et soit $r_\beta(h)$ la longueur critique pour qu'un cluster de poids plus grand que h soit significatif au niveau β i.e.

$$\mathbb{P}(L_h \leq r_\beta(h)) = \beta.$$

L'idée du test est de trouver le "bon" niveau β et, pour chaque poids h possible, la longueur critique correspondante $r_\beta(h)$, de telle façon que

1.
$$\mathbb{P}(L_h \leq r_\beta(h)) = \beta;$$

2.
$$\mathbb{P}(\bigcup_h \{L_h \leq r_\beta(h)\}) = \alpha,$$

où α est l'erreur de première espèce du test.

Ayant toutes les longueurs critiques pour tous les poids possibles h, on peut décider de la significativité des clusters d'orthologues observés en comparant leur longueur à la longueur critique correspondante à leur poids.

Pour un h et un r donnés, la probabilité $\mathbb{P}(L_h \leq r)$ est simplement la probabilité de trouver quelque part dans le génome B une région de poids plus grand que h et de longueur plus petite que r, i.e. $\mathbb{P}(W_m \geq 1)$ avec W_m défini comme précédemment. On calcule donc cette probabilité à l'aide de l'approximation de Poisson composée décrite ci-dessus.

La solution à laquelle nous avons pensé pour trouver le seuil β et les longueurs critiques $r_\beta(h), \forall h$ vérifiant les conditions 1 et 2 est la suivante.

Pour un β donné, on trouve les longueurs critiques $r_\beta(h), \forall h$ qui vérifient la condition 2 en utilisant une méthode numérique de résolution d'équations. Ensuite, on calcule par la méthode de Monte Carlo la probabilité de la réunion sur les h, qui va être vue comme une fonction de β. Pour finalement trouver le "bon" β tel que la condition 1 soit vérifiée, on utilise encore une fois une méthode numérique de résolution d'équations.

Par contre, pour l'utilisation pratique de ce test, nous avons rencontré des grandes difficultés techniques liées à la complexité des calculs.

Nous espérons trouver dans un travail futur une solution plus convenable à ce problème.

Mesures pour l'exceptionalité de l'ordre des gènes

Dans la deuxième partie de notre travail de thèse nous nous sommes intéressés à l'ordre des gènes dans des régions génomiques conservées, avec l'idée que les clusters dans lesquels l'ordre des orthologues est plus conservé sont encore plus significatifs du point de vue biologique.

Un des problèmes rencontrés a été de trouver une "bonne" mesure pour quantifier le degré de conservation des orthologues dans les clusters. Ici, "bonne" signifie biologiquement pertinente et en même temps accessible du point de vue des calculs.

Dans le Chapitre 2 nous proposons trois mesures pour quantifier l'exceptionalité de l'ordre des gènes dans des régions génomiques conservées trouvées par une approche de type région de référence.

Nous traitons seulement le cas sans famille multigénique, i.e. nous supposons que pour chaque gène de la région de référence on trouve au plus un orthologue dans le génome B.

Le cas avec des familles multigéniques revient à la comparaison de multipermutations et est beaucoup plus difficile à traiter. Nous n'avons pas encore trouvé une bonne solution à ce problème et il reste un problème ouvert pour des travaux futurs. C'est la raison pour laquelle nous n'avons pas testé les résultats de ce chapitre sur des données réelles, qui contiennent des co-orthologues.

Les trois mesures que nous présentons sont basées sur la distance de transposition dans le groupe des permutations. Nous obtenons des expressions analytiques pour leur

distribution dans le cas d'une permutation aléatoire, i.e. sous l'hypothèse nulle d'ordre aléatoire des gènes.

Pour comparer deux permutations dans le groupe symétrique S_n on utilise la distance de transposition dans le groupe symétrique S_n, que l'on va noter d_t.

Au premier regard, la distance de transposition dans le groupe des permutations ne semble pas très pertinente du point de vue biologique, car les transpositions mathématiques ne correspondent à aucun évènement génomique.

Dans la littérature sur les réarrangements génomiques, plusieurs distances plus pertinentes biologiquement ont été étudiées, distances qui prennent en compte un ou une combinaison de réarrangements génomiques : inversions, translocations, fissions et fusions chromosomiques, transpositions biologiques, échanges de blocs ("block-interchanges" en anglais) – voir [27].

Le problème avec l'utilisation de ces distances comme statistiques de test provient du fait que leur distribution, pour une permutation aléatoire, est très difficile à obtenir. Dans la littérature il y a très peu de résultats sur ce sujet. Récemment, Doignon et Labarre [13] ont trouvé la distribution du nombre de cycles alternés dans le graphe des points de rupture ("breakpoint graph" en anglais) d'une permutation (non-signée) aléatoire, résultat qui peut être utilisé pour déduire la distribution de ces distances génomiques qui sont basées sur le graphe des points de rupture, comme la distance d'échange de blocs ("block-interchange distance" en anglais) de Christie [9]. Sankoff et Haque [34] et Xu, Zheng et Sankoff [35] ont utilisé une approche constructive pour obtenir des estimations asymptotiques pour la distribution du nombre de cycles dans le graphe de points de rupture de deux permutations signées aléatoires (pour la définition du graphe des points de rupture, voir la Section 3.1).

Dans notre travail nous utilisons la distance de transposition parce qu'elle est très convenable du point de vue des calculs.

Mais on pourra se demander si cette distance peut avoir du sens du point de vue biologique.

Une réponse positive à cette question est donnée par Eriksen et Hultman dans [15], où ils décrivent une analogie entre les transpositions mathématiques et les inversions génomiques. Ils montrent que la distance moyenne de transposition après t transpositions aléatoires appliquées à la permutation identique est une bonne approximation pour la distance moyenne d'inversion, pour un génome avec n gènes, après t inversions aléatoires appliquées à l'identité. Après avoir obtenu une formule explicite pour la première, ils proposent une méthode pour estimer la vraie distance évolutive entre deux génomes et ils montrent que cette méthode se comporte très bien en comparaison avec les meilleurs résultats obtenus par d'autres méthodes.

L'originalité des mesures présentées dans les Sections 2.4 et 2.5 repose sur le fait qu'elles ne prennent pas en compte seulement l'ordre des gènes orthologues qui sont en commun entre les deux clusters, mais aussi les positions des autres orthologues. Ces mesures sont spécifiquement adaptées au cas des clusters trouvés par une approche de type région de référence.

Nous présentons dans la suite les trois mesures que nous avons utilisées, ainsi que les résultats que nous avons obtenus sur leur distribution sous l'hypothèse nulle.

Soit n le nombre de gènes dans la région de référence ayant un et un seul orthologue dans le génome B.

On étiquette les n orthologues de telle sorte que leur ordre dans la région de référence

soit la permutation idéntique Id_n, et on note π la permutation qui représente l'ordre des orthologues dans le génome B.

Supposons que nous sommes intéressés à l'ordre des gènes dans une certaine région génomique conservée $\mathcal{R}$ trouvée dans le génome B, contenant h orthologues et commençant avec le i-ème orthologue, étiquetté $\pi(i)$.

Sous l'hypothèse nulle H_0, π est une permutation aléatoire dans S_n, choisie uniformément avec probabilité $1/n!$.

Pour une permutation $\pi \in S_n$, on va utiliser la notation $\pi = [\pi(1), ..., \pi(n)]$.

Notation. *Pour $\sigma \in S_n$ et $j \in \{1, ..., n-h+1\}$, on note $\sigma_{j,h}$ la restriction de σ à l'ensemble $\{j, j+1, ..., j+h-1\}$, i.e. $\sigma_{j,h} = [\sigma(j), ..., \sigma(j+h-1)]$.*

Une première distance

Une première idée a été de comparer seulement l'ordre des h orthologues dans la région $\mathcal{R}$, donnée par $\pi_{i,h}$, avec leur ordre dans la région de référence, donnée par $Id_n|_{\{\pi(i),...,\pi(i+h-1)\}}$.

On utilise la notation suivante.

Notation. *Pour $1 \leq k \leq n$, on note*

$$p(n,k) := \frac{1}{n!} \sum_{j=k}^{n} s(n,j)$$

où $s(n,j)$ sont des nombres de Stirling non-signés.

En utilisant des résultats classiques sur les permutations (voir les Lemmes 2.1 et 2.2), on remarque que $p(n,k)$ représente la probabilité qu'une permutation aléatoire dans S_n ait au moins k cycles.

On remarque que la loi de $d_t(\pi_{i,h}, Id_n|_{\{\pi(i),...,\pi(i+h-1)\}})$ pour une permutation aléatoire π dans S_n est la même que la loi de la distance de transposition pour une permutation aléatoire dans S_h, donc on obtient le résultat suivant.

Proposition (Proposition 2.3)**.** *Pour $0 \leq d \leq h-1$, on a*

$$\mathbb{P}(d_t(\pi_{i,h}, Id_n|_{\{\pi(i),...,\pi(i+h-1)\}}) \leq d) = p(h, h-d).$$

Une deuxième distance

Avec la première distance on ne considère que l'ordre relatif des h gènes qui sont en commun entre la région $\mathcal{R}$ et la région de référence, en ignorant l'ordre des autres orthologues. Par exemple, si $n = 20$ et $h = 5$, avec cette première distance on va donner le même poids à une région ayant $\pi_{i,h} = [1,2,3,4,5]$ et à une région ayant $\pi_{i,h} = [1,5,10,15,20]$. Par contre, on aura envie de dire que la première région est plus significative, du point de vue de l'ordre des gènes, que la deuxième.

Une deuxième idée sera de prendre en compte aussi les positions, dans la région de référence, des autres $n-h$ orthologues, mais d'ignorer toujours leur ordre dans le génome B.

La deuxième "distance" considérée est la suivante :

$$d(\pi_{i,h}, Id_n) := \min\{d_t(\sigma, Id_n) : \sigma \in S_n, \sigma_{i,h} = \pi_{i,h}\},$$

i.e. on prend le minimum sur toutes les possibilités d'ordonner dans B les $n-h$ orthologues qui ne se trouvent pas dans la région $\mathcal{R}$.

Cette distance tient compte de la position dans B de la région $\mathcal{R}$.

Nous avons obtenu explicitement la loi de $d(\pi_{i,h}, Id_n)$ sous l'hypothèse nulle, i.e. pour une permutation aléatoire π dans S_n.

Théorème (Theorem 2.5)**.** *Pour $0 \leq d \leq h-1$, on a*

$$\mathbb{P}(d(\pi_{i,h}, Id_n) \leq d) = \frac{1}{\binom{n}{h}} \sum_{m=h-d}^{h} \binom{n-m-1}{n-h-1} p(m, h-d).$$

Du point de vue biologique, cette distance paraît un peu trop restrictive par rapport à la position du cluster $\mathcal{R}$ dans le génome B.

Pour prendre en compte des évènements génomiques éventuels (comme des translocations ou des transpositions génomiques) qui ont pu changer la position dans B de la région $\mathcal{R}$ par rapport aux autres orthologues, nous avons pensé utiliser la distance suivante :

$$d_{min}(\pi_{i,h}, Id) := \min_{1 \leq k \leq n-h+1} \min\{d_t(\sigma, Id) : \sigma \in S_n, \sigma_{k,h} = \pi_{i,h}\}.$$

Par contre, nous n'avons pas réussi à trouver la loi de cette distance pour une permutation aléatoire.

Une troisième distance

Une solution de compromis que nous avons trouvée est d'utiliser la "distance" suivante :

$$d^*(\pi_{i,h}, Id_n) := \min\{d_t(\sigma, Id_n) : \sigma \in S_n, \sigma_{i^*,h} = \pi_{i,h}\},$$

où

$$i^* := \arg \max_{1 \leq j \leq n-h+1} \left|\{\pi_i, ..., \pi_{i+h-1}\} \cap \{j, ..., j+h-1\}\right|.$$

Par convention, dans le cas où il y a plusieurs points de maximum, on décide de choisir i^* comme le plus petit parmi eux.

Pour trouver la loi de cette distance, nous avons conditionné par

$$L^* := \left|\{\pi_i, ..., \pi_{i+h-1}\} \cap \{i^*, ..., i^*+h-1\}\right|,$$

qui est une statistique de scan discrète conditionnelle dont la loi exacte est donnée par la Proposition 2.7.

Nous avons obtenu le résultat suivant pour la loi de la distance d^*.

Théorème (Theorem 2.8)**.** *Pour $0 \leq d \leq h-1$, on a*

$$\mathbb{P}(d^*(\pi_{i,h}, Id_n) \leq d) = \frac{1}{\binom{h}{\ell}} \sum_{\ell=h-d}^{h} \mathbb{P}(L^* = \ell) \sum_{m=h-d}^{\ell} \binom{h-m-1}{h-\ell-1} p(m, h-d),$$

où

$$\mathbb{P}(L^* = \ell) = \mathbb{P}(N_h < \ell+1 | X_{1,n} = h) - \mathbb{P}(N_h < \ell | X_{1,n} = h)$$

et les deux probabilités conditionnelles sont données par la Proposition 2.7.

Nos résultats peuvent aider à renforcer la puissance des tests de significativité pour des régions génomiques conservées, qui ne prendraient en compte que la proximité des gènes et pas leur ordre. Par contre, comme Sankoff et Haque [33] l'ont déjà remarqué, comment combiner, dans un seul test statistique, la p-valeur basée sur la proximité des orthologues et celle basée sur leur ordre n'est pas du tout clair.

Supposons que nous sommes intéressés par la significativité d'un certain cluster qui contient h orthologues, commence avec le i-ème orthologue, est de longueur r et pour lequel la distance d^* vaut d.

Une idée que nous avons eu est de pondérer l'importance de l'ordre des orthologues dans le cluster en introduisant un paramètre supplémentaire $\gamma \in [0,1]$ et en prenant comme p-valeur combinée le produit

$$p(\gamma) := \mathbb{P}(L_h \leq r)\mathbb{P}(d^*(\pi_{i,h}, Id_n) \leq d)^\gamma.$$

On pourra tracer la courbe de $p(\gamma)$ quand γ varie dans $[0,1]$. Par contre, le choix du paramètre γ n'est pas évident.

Distribution du nombre de cycles dans le "breakpoint graph"

Dans le Chapitre 3 nous nous intéressons à la distribution du nombre des cycles dans le graphe des points de rupture ("breakpoint graph") d'une permutation signée aléatoire.

La connaissance de cette distribution fournit par la suite une très bonne approximation pour la distribution de la distance d'inversion pour une permutation signée aléatoire.

En effet, Bafna et Pevzner [3] ont donné la borne inférieure suivante pour la distance d'inversion ("reversal distance" en anglais) :

$$d_{rev}(\pi, Id) \geq n + 1 - c(\pi),$$

où π est une permutation signée de n éléments et $c(\pi)$ représente le nombre de cycles alternés dans le graphe des points de rupture de π. Cette borne inférieure approxime très bien la distance d'inversion, tant pour des données simulées (voir Kececioglu et Sankoff [26]), que pour des données biologiques (voir Bafna et Pevzner [3]).

En utilisant la technique "Markov chain imbedding", introduite par Fu et Koutras [16], nous obtenons la distribution du nombre de cycles dans le graphe des points de rupture d'une permutation signée aléatoire, sous la forme d'un produit de matrices de transition d'une certaine chaîne de Markov finie.

Dans la sous-section 3.2.1 nous donnons une courte présentation de la méthode "Markov chain imbedding". Ensuite, dans la sous-section 3.2.2, nous appliquons cette méthode à notre problème. Nous décrivons la chaîne de Markov finie, non-homogène, qui va nous permettre de retrouver la distribution du nombre de cycles dans le graphe des points de rupture, et nous obtenons les matrices de transition de cette chaîne (voir la Proposition 3.1).

Dans la suite nous présentons brièvement notre approche.

On commence par définir le graphe des points de rupture d'une permutation signée aléatoire.

A chaque permutation signée π de n éléments on associe une permutation (non-signée) $\pi' \in S_{2n}$ en replaçant tout élément positif $+i$ par la paire $(2i-1, 2i)$ et tout élément négatif

$-i$ par la paire $(2i, 2i-1)$. Ensuite, on étend π' en lui ajoutant deux nouveaux éléments, un au debut, noté S (de "Start"), et un à la fin, noté T (de "Terminus").

Le graphe des points de rupture ("breakpoint graph") d'une permutation signée π de n éléments est le graphe $G(\pi) = (V, B \cup C)$ ayant l'ensemble des sommets $V = \{S, 1, 2, ..., 2n, T\}$ et l'ensemble des arêtes partitionné en deux : le sous-ensembles B des arêtes *pleines*, correspondant aux adjacences dans la permutation π, et l'ensemble C des arêtes *en tirets*, correspondant aux adjacences dans la permutation identique Id (voir la Définition 3.1 et Fig. 3.1).

Le graphe $G(\pi)$ se décompose de façon unique dans des cycles alternés, i.e. des cycles dans lesquels les arêtes pleines et les arêtes en tirets alternent. On note $c(\pi)$ le nombre de cycles alternés dans le graphe $G(\pi)$.

Pour un cycle donné, sa *longueur* est le nombre d'arêtes pleines qu'il contient.

Nous construisons une chaîne de Markov finie de la façon suivante.

Soit π_1 une permutation signée aléatoire de 1 élément, i.e. $\pi_1 = [+1]$ avec probabilité $1/2$ et $\pi_1 = [-1]$ avec probabilité $1/2$.

Pour chaque $t = 2, ..., n$, soit π_t la permutation signée aléatoire de t éléments qui est obtenue à partir de π_{t-1} en insérant au hasard l'élément t, uniformément, dans une des t positions possibles, avec le signe "+" avec probabilité $1/2$ et le signe "-" avec probabilité $1/2$, le signe étant indépendant de la position.

Sous l'hypothèse nulle, π_n est une permutation signée aléatoire de n éléments. On veut trouver la loi de $c(\pi_n)$.

Pour chaque $t = 1, ..., n$, on note $K_{j,t}, j = 1, ..., n+1$ le nombre de cycles de longueur j dans $G(\pi_t)$. On note aussi L_t la longueur du cycle de $G(\pi_t)$ qui contient le point terminal T.

La chaîne de Markov finie qui va nous permettre de retrouver la loi de $c(\pi_n)$ est la suivante :

$$Y_t := (L_t, K_{1,t}, ..., K_{n+1,t}), t = 1, ..., n.$$

En effet, nous avons

$$\mathbb{P}(c(\pi_n) = x) = \mathbb{P}(Y_n \in C_x),$$

où, pour chaque $x = 1, 2, ..., n+1$:

$$C_x = \left\{ (\ell, k_1, ..., k_{n+1}) : \sum_{j=1}^{n+1} k_j = x, \sum_{j=1}^{n+1} j k_j = n+1, k_\ell \geq 1 \right\}.$$

Nous avons le résultat suivant.

Proposition (Proposition 3.1)**.** *$(Y_t)_{1 \leq t \leq n}$ est une chaîne de Markov non-homogène de loi initiale*

$$\mathbb{P}(Y_1 = (1, 2, 0, 0, ..., 0)) = \mathbb{P}(Y_1 = (2, 0, 1, 0, ..., 0)) = 1/2,$$

et de probabilités de transition suivantes. Si $Y_{t-1} = (\ell, \overrightarrow{k})$, avec $k_\ell \geq 1$, alors les transitions possibles sont vers $Y_t = (\ell', \overrightarrow{k'})$, où

(i) $\ell' = \ell + 1$ and $\overrightarrow{k'} = \overrightarrow{k} - e'_\ell + e'_{\ell+1}$, avec probabilité $\ell/(2t)$;

(ii) $\ell' = j$, avec $1 \leq j \leq \ell$ et $\overrightarrow{k'} = \overrightarrow{k} - e'_\ell + e'_j + e'_{\ell+1-j}$, avec probabilité $1/(2t)$;

(*iii*) $\ell' = \ell + x + 1$, *avec* $1 \leq x \leq t - \ell, x \neq \ell$ *et* $\overrightarrow{k'} = \overrightarrow{k} - e'_\ell - e'_x + e'_{\ell+x+1}$, *avec probabilité* xk_x/t;

(*iv*) $\ell' = 2\ell + 1$ *et* $\overrightarrow{k'} = \overrightarrow{k} - 2e'_\ell + e'_{2\ell+1}$, *avec probabilité* $\ell(k_\ell - 1)/t$,

où pour chaque i, e'_i *est le vecteur ligne ayant* 1 *à la* i*-ème position et* 0 *ailleurs.*

Un point faible de notre approche est le fait que la dimension de la chaîne de Markov est très grande, ce qui entraîne une grande complexité algorithmique. Nous avons implémenté en Matlab une procédure itérative qui calcule, pour un n donné, la distribution du nombre de cycles dans le graphe des points de rupture d'une permutation signée aléatoire avec n éléments.

Par contre, notre méthode nous permet d'obtenir des relations de récurrence pour la distribution du nombre de cycles dans le graphe des points de rupture, que nous espérons arriver à résoudre explicitement dans un travail futur.

Une application de nos résultats à la détection de régions génomiques conservées est d'utiliser la distance d'inversion comme mesure de l'exceptionalité de l'ordre des gènes orthologues dans des clusters de gènes, dans le cas des génomes signés, i.e. pour lesquels on connaît aussi l'orientation des gènes. On pourra utiliser cette distance à la place de la distance de transposition, avec l'idée de la Section 2.3, où on ne compare que l'ordre relatif des orthologues qui sont en commun entre la région conservée dans B qui nous intéresse et la région de référence.

Applications à des données biologiques

Le Chapitre 4 est dédié à deux applications des résultats du Chapitre 1 sur des données biologiques.

Le premier jeu de données correspond à une comparaison entre une région du Complexe Majeur d'Histocompatibilité du génome humain et le génome d'un poisson, Oryzias-Latipes.

Le deuxième jeu de données correspond à une comparaison entre une région de Ciona-Intestinalis et le génome humain.

Chapitre 1

Compound Poisson approximation and testing for conserved genomic regions

1.1 Biological context and related work

Orthologous genes are two genes, in two different species, that descend from the same gene at the ancestor of the two species, as the result of a speciation event. They tend, in general, to have similar functions. Therefore, a group of genes that cluster together in two different species in a way that is significantly improbable by chance, may be either the mark of evolutionary relationships between the two species, or the sign of a functional selective pressure acting on these genes.

We call *conserved genomic region* or *gene cluster* two chromosomic regions, in two different species, that have in common a certain number of orthologous genes, not necessarily adjacent or in the same order in the two genomes. We do not impose any restriction on the gap length between consecutive orthologs.

The conserved genomic regions play an important role in the attempt to reconstruct the evolutionary history of the species, and can also serve to infer functional relationships between genes.

In the literature various definitions for gene clusters exist (see [6, 11, 14, 22, 23, 29]). We have chosen here a very unrestrictive definition, in order to be able to detect evolutionary signals even between very distant species.

During the evolutionary time, the gene order in one genome can be affected by various genome rearrangement events, like inversions, translocations, transpositions, chromosomic fissions and fusions. Hence, in the absence of certain constraints due to functional selective pressures, the gene order is rapidly randomized. This is one reason why, in general, the null hypothesis taken in the significance tests for gene clusters is the hypothesis of random gene order.

There are different approaches when searching for gene clusters (see [14]). In this work we focus on the case when the gene clusters are found by the "reference region" approach, which consists in starting with a fixed genomic region in a certain species A (called the *reference region*) and searching for significant orthologous gene clusters in the genome of

another species B.

In general, the orthology relation between the genes of two species is not one-to-one. For a given gene in one species we may find more than one orthologous gene in another species, as the result of duplication events happened after the separation of the two species. The genes in one species which are orthologous to the same gene in another species are called *co-orthologs* of this gene and form what we call a *multigene family*.

The existence of multigene families is an important fact which needs to be considered when testing for gene cluster significance, but very few of the existent statistical tests consider it. Danchin and Pontarotti [11] propose to weight the orthologs in inverse proportion to the sizes of the multigene families, but their use of a binomial distribution is not adequate in these settings. Raghupathy and Durand [29] take also into account the existence of multigene families, but their test is suitable only for clusters found by the window-sampling approach, and not by the reference region approach, as in our case.

In this work we adopt the idea of Danchin and Pontarotti [11] for taking into account the multigene families and propose a compound Poisson approximation for computing the probabilities, under the null hypothesis, of different gene clusters.

1.2 Mathematical framework

1.2.1 Mathematical formulation of the problem

We model the genome as an ordered set of genes, the length of a genomic region being measured in number of genes. We ignore the separation into chromosomes and the physical distances between genes.

The data that we dispose of are : the number m of genes in the reference region from the genome A which have at least one ortholog in the genome B ; for each of those genes $i = 1, ..., m$, the number of orthologs it has in B, which we denote ϕ_i ; the positions in B of these orthologs ; the total number N of genes in the genome B.

We make the (natural) assumption that there exists a maximal size $\phi_{\max}$ for the multigene families.

Based on the fact that we are in the case $m << N$, we make a further approximation and consider the genome B as the continuous interval $[0, 1]$, in which the "new" positions of the orthologs are obtained by dividing by N their real positions in the genome.

We will use a pure significance test, with the null hypothesis being the hypothesis

$$H_0 : \text{ random gene order in the genome B.}$$

All the probabilities and distributions appearing throughout the paper are implicitly considered under the null hypothesis H_0.

For $i = 1, ..., m$ we let $U_{ij}, j = 1, ..., \phi_i$ represent the positions in B of the orthologs of the gene i from the reference region. Under H_0, the r.v.'s $U_{ij}, j = 1, ..., \phi_i, i = 1, ..., m$ are i.i.d. uniformly distributed on $[0, 1]$.

Let $n := \phi_1 + \cdots + \phi_m$ denote the total number of genes in B which are orthologous to genes in the reference region in A.

We are interested only in these n genes. We want to test whether they cluster together in a significant way, i.e. in a way which is very improbable by chance, under the null hypothesis.

For taking into account the existence in B of multiple orthologs for the genes in the reference region, we consider the following counting measure :

$$\mu_m := \sum_{i=1}^{m} \frac{1}{\phi_i} \sum_{j=1}^{\phi_i} \delta_{U_{ij}}.$$

For an ortholog belonging to a multigene family of size ϕ_i, we call $\frac{1}{\phi_i}$ its *label*. For an interval $I \subset [0,1]$, we will refer to $\mu_m(I)$ as its *weight*.

In the simple case of no multigene families (all ϕ_i's equal to one) $\mu_m(I)$ is just the number of orthologs lying in the interval I.

For applying a statistical test we need to compute, for a given weight h and a given length r, the probability, under the null hypothesis, of finding somewhere in the genome B an orthologous cluster of weight greater than h and of length smaller than r. We will call such a cluster *of type* $(h : r)$.

For technical simplifications, we first consider the case when B is a circular genome, hence the circle of length 1 in our model.

1.2.2 The circular case

Let h be fixed, of the form

$$h = \sum_{i=1}^{m} \frac{n_i}{\phi_i}, \text{ with } 0 \leq n_i \leq \phi_i, i = 1, ..., m.$$

Let $r \in (0,1)$ be also fixed.

We denote by $U_{(1)} \leq U_{(2)} \leq \cdots \leq U_{(n)}$ the ordered positions in B of the n orthologs, i.e. the order statistics of n i.i.d. r.v.'s uniformly distributed on the circle of length 1.

Let $W_m = W_m(h,r)$ denote the r.v. representing the number of (possibly overlapping) clusters of type $(h : r)$ in the genome B.

Let also

$$A_k = A_k(h,r) = \{\mu_m([U_{(k)}, U_{(k)} + r]) \geq h\}$$

denote the event of having in B a cluster of type $(h : r)$ starting with the k-th ortholog.

We have

$$W_m = \sum_{k=1}^{n} 1_{A_k}.$$

Note that

$$\mathbb{P}(W_m \geq 1) = \mathbb{P}(\bigcup_{k=1}^{n} A_k)$$

is the probability of finding, somewhere in the genome B, at least one cluster of type $(h : r)$.

We are interested in computing this probability.

We will further simplify the parameterization of the problem.

Let $\phi'_1 < \cdots < \phi'_J$ denote all the different values among the multigene families' sizes $\phi_1, ..., \phi_m$, and let $g_j = \left|\{i = 1, ..., m : \phi_i = \phi'_j\}\right|, j = 1, ..., J$ denote their multiplicities.

Therefore, we have

$$\begin{cases} \sum_{j=1}^{J} g_j = m, \\ \sum_{j=1}^{J} g_j \phi'_j = n. \end{cases}$$

Remark. *We can represent the measure μ_m as*

$$\mu_m = \sum_{i=1}^{n} L_i \delta_{U_{(i)}},$$

where δ_x denotes the Dirac measure in x and $\boldsymbol{L} = (L_1, ..., L_n)$ is a random vector independent of the $U_{(i)}$'s and uniformly distributed over the set

$$\Lambda = \left\{ \boldsymbol{\ell} = (\ell_1, ..., \ell_n) \in \left\{ \frac{1}{\phi'_1}, ..., \frac{1}{\phi'_J} \right\}^n : \left| \left\{ i : \ell_i = \frac{1}{\phi'_j} \right\} \right| = g_j \phi'_j, \forall j \right\}$$

of all possible labellings of the n orthologs.

Note that Λ is the set of all the permutations of a multiset with multiplicities $g_j \phi'_j$, $j = 1, ..., J$, and so :

$$|\Lambda| = \frac{n!}{(g_1 \phi'_1)! \cdots (g_J \phi'_J)!}.$$

Let $n_{\min} := g_1 = |\{i : \phi_i = \phi'_1\}|$, where $\phi'_1 = \min\{\phi_i : i = 1, ..., m\}$. We let also $h_* := \lceil h\phi'_1 \rceil$ and we assume that $n_{\min} \geq h_*$, s.t. h_* is the minimal number of orthologs in a cluster of weight greater than h.

For every labelling $\boldsymbol{\ell} \in \Lambda$ and for every $k = 1, ..., n$, let

$$h_k(\boldsymbol{\ell}) := \min\{d : \ell_k + \cdots + \ell_{k+d-1} \geq h\}$$

be the minimal number of orthologs in a cluster starting with the k-th ortholog so as to be of weight greater than h.

Therefore,

$$A_k \cap \{\boldsymbol{L} = \boldsymbol{\ell}\} = \{U_{(k+h_k(\boldsymbol{\ell})-1)} - U_{(k)} \leq r\} \cap \{\boldsymbol{L} = \boldsymbol{\ell}\}.$$

Let also

$$h^* := \max_{\boldsymbol{\ell}} \{h_1(\boldsymbol{\ell})\}.$$

We have $h^* \leq \lceil h\phi'_J \rceil$, where $\phi'_J = \max\{\phi_i : i = 1, ..., m\}$.

Remark. *We place ourselves in the asymptotic settings of $m \longrightarrow \infty$ or, equivalently, $n \longrightarrow \infty$.*

Note that we are in the case of a sum of indicators which are in a *short-range dependence* and a *long-range (almost) independence*. Because of the strong dependence between the neighbouring indicators, the events A_k will tend to occur in clumps. Consequently, it seems reasonable to approach the distribution of W_m by a compound Poisson distribution, with a "good" choice of the parameters.

1.3 The Stein-Chen method for compound Poisson approximation

Definition 1.1. *Let* $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ...)$ *be s.t.* $|\boldsymbol{\lambda}| := \sum_{i=1}^{\infty} \lambda_i < \infty$. *The* Compound Poisson distribution of parameter $\boldsymbol{\lambda}$ *is*

$$CP(\boldsymbol{\lambda}) := \mathcal{L}(\sum_{i=1}^{\infty} iZ_i) = \mathcal{L}(\sum_{j=1}^{M} X_j),$$

where

- $Z_i \sim Poisson(\lambda_i)$, *independent;*
- $(X_j)_j$ *i.i.d. with distribution* $\frac{1}{|\boldsymbol{\lambda}|}\boldsymbol{\lambda}$, *independent of* $M \sim Poisson(|\boldsymbol{\lambda}|)$.

In our case, the variables appearing in this definition have the following interpretation :
– Z_i represents the number of clumps containing i consecutive events. The assumption that the Z_i's are independent and Poisson distributed is based on the long-range independence between the events A_k.
– M represents the total number of clumps. For each j, X_j represents the size of the j-th clump (i.e. the number of consecutive events it contains). The assumption that M has a Poisson distribution is also related to the long-range independence between the events.

Note that λ_i represents the mean number of clumps of size i.

We will use only compound Poisson distributions of finite expectation, i.e. verifying $\sum_{i=1}^{\infty} i\lambda_i < \infty$.

We will approximate the distribution of W_m by a compound Poisson distribution $CP(\boldsymbol{\lambda})$ and we will quantify the error using the Kolmogorov distance.

Definition 1.2. *The Kolmogorov distance between two distributions* μ *et* ν *on* $\mathbb{R}_+$ *is*

$$d_K(\mu, \nu) = \sup_{k \in \mathbb{N}} |\mu([k, \infty) - \nu([k, \infty))| .$$

We will approximate our probability of interest $\mathbb{P}(W_m \geq 1)$ by the corresponding probability for the compound Poisson distribution, $CP(\boldsymbol{\lambda})([1, \infty)) = 1 - \exp\{-\sum_{i=1}^{\infty} \lambda_i\}$, with an error

$$|\mathbb{P}(W_m \geq 1) - CP(\boldsymbol{\lambda})([1, \infty))| \leq d_K(\mathcal{L}(W_m), CP(\boldsymbol{\lambda})).$$

Therefore, it suffices to obtain bounds for the Kolmogorov distance between the two distributions.

For bounding the Kolmogorov distance we will use the Stein-Chen method for compound Poisson approximation, introduced by Barbour, Chen and Loh [4].

In the following we will briefly present this method.

1.3.1 The Stein-Chen method

Let W be a r.v. that we want to approximate by a compound Poisson distribution. The Stein-Chen method gives a way to measure the precision of the approximation.

Definition 1.3. *The* Stein-Chen equation *is the following :*

$$jg(j) - \sum_{i=1}^{\infty} i\lambda_i g(j+i) = f_{\lambda,A}(j) := \mathbf{1}_{\{j\in A\}} - CP(\boldsymbol{\lambda})(A),\ j \geq 0,$$

for every $A \subset \mathbb{N}$, the unknown being the function $g : \mathbb{N} \longrightarrow \mathbb{R}$.

Barbour, Chen and Loh [4] proved that this equation has a bounded solution g which is unique, except for the value in $g(0)$. But the non-unicity in 0 is not a problem in our case, because the value at 0 does not appear in the computations.

Remark. *We have the following properties :*

(a) $\mathbb{E}f_{\lambda,A}(Z) = \mathbb{P}(Z \in A) - CP(\boldsymbol{\lambda})(A)$ *for every r.v. Z.*

(b) If $Z \sim CP(\boldsymbol{\lambda})$*, then* $\mathbb{E}f_{\boldsymbol{\lambda},A}(Z) = 0$.

(c) If $g : \mathbb{N} \longrightarrow \mathbb{R}$ *is bounded and* $Z \sim CP(\boldsymbol{\lambda})$ *with* $\sum_i i\lambda_i < \infty$*, then*

$$\mathbb{E}\left\{Zg(Z) - \sum_{i=1}^{\infty} i\lambda_i g(Z+i)\right\} = 0.$$

If $g_{\boldsymbol{\lambda},A}$ is the solution of the Stein-Chen equation, by evaluating at $j = Z$ and taking the expectation, we obtain

$$\mathbb{E}\left\{Wg_{\lambda,A}(W) - \sum_{i=1}^{\infty} i\lambda_i g_{\lambda,A}(W+i)\right\} = \mathbb{P}(W \in A) - CP(\boldsymbol{\lambda})(A).$$

Consequently, if $\mathcal{F}$ is a test set of events, we deduce the following expression for the corresponding distance :

$$\begin{aligned} d_{\mathcal{F}}(\mathcal{L}(W), CP(\boldsymbol{\lambda})) &:= \sup_{A\in\mathcal{F}} |\mathbb{P}(W \in A) - CP(\boldsymbol{\lambda})(A)| = \sup_{A\in\mathcal{F}} |\mathbb{E}f_{\boldsymbol{\lambda},A}(W)| \\ &= \sup_{A\in\mathcal{F}} \left|\mathbb{E}\left\{Wg_{\boldsymbol{\lambda},A}(W) - \sum_{i=1}^{\infty} i\lambda_i g_{\boldsymbol{\lambda},A}(W+i)\right\}\right|. \end{aligned}$$

The idea of the method is to show that

$$\left|\mathbb{E}\left\{Wg(W) - \sum_{i=1}^{\infty} i\lambda_i g(W+i)\right\}\right| \leq ||\Delta g||\varepsilon \tag{1.1}$$

holds for every bounded function $g : \mathbb{N} \longrightarrow \mathbb{R}$, where

$$||\Delta g|| := \sup_{i\geq 1} |g(i+1) - g(i)|.$$

This will further imply that

$$d_{\mathcal{F}}(\mathcal{L}(W), CP(\boldsymbol{\lambda})) \leq c_{\mathcal{F}}(\boldsymbol{\lambda})\varepsilon, \tag{1.2}$$

where

$$c_{\mathcal{F}}(\boldsymbol{\lambda}) := \sup_{A\in\mathcal{F}} ||\Delta g_{\boldsymbol{\lambda},A}||.$$

Barbour and Xia [5] showed that for $\mathcal{F} = \{[k, \infty) : k \in \mathbb{N}\}$ (corresponding to the Kolmogorov distance), if the following condition is fulfilled :

$$\lambda_1 \geq 2\lambda_2 \geq 3\lambda_3 \geq \cdots, \tag{1.3}$$

then we have

$$c_K(\boldsymbol{\lambda}) := c_{\mathcal{F}}(\boldsymbol{\lambda}) \leq \min\left\{\frac{1}{2}, \frac{1}{\lambda_1 + 1}\right\}. \tag{1.4}$$

1.3.2 The coupling approach

There are two approaches in the Stein-Chen method, one called the "local approach" and one called the "coupling approach" (see Roos [30, 31]).

Here we use the coupling approach, which consists in finding the upper bound appearing in (1.1) by means of a certain coupling which we will describe in the sequel.

Let $W = \sum_{\alpha\in\Gamma} I_\alpha$ be a sum of indicators. We assume that there exists a local-dependence structure between these indicators (of the type short-range dependence, long-range independence), so that we can, for every $\alpha \in \Gamma$, divide Γ into four disjoint subsets $\{\alpha\}, \Gamma_\alpha^{vs}, \Gamma_\alpha^{vw}$ and Γ_α^b :

$$\begin{aligned}
\Gamma_\alpha^{vs} &:= \{\beta \in \Gamma\backslash\{\alpha\} : I_\beta \text{ "very strongly" dependent on } I_\alpha\},\\
\Gamma_\alpha^{vw} &:= \{\beta \in \Gamma\backslash\{\alpha\} : I_\beta \text{ "very weakly" dependent on } \{I_\gamma, \gamma \in \{\alpha\} \cup \Gamma_\alpha^{vs}\}\},\\
\Gamma_\alpha^{b} &:= \Gamma\backslash\{\{\alpha\} \cup \Gamma_\alpha^{vs} \cup \Gamma_\alpha^{vw}\}.
\end{aligned}$$

We let

$$U_\alpha := \sum_{\beta\in\Gamma_\alpha^{vs}} I_\beta,\ Z_\alpha := I_\alpha + U_\alpha,\ X_\alpha := \sum_{\beta\in\Gamma_\alpha^{b}} I_\beta.$$

For every $\alpha \in \Gamma$, let V_α be a r.v. and $\mathcal{V}_\alpha$ its set of values.

We have the following theorem of Roos (Theorem 4.G. in [30]).

Theorem 1.1. *Assume that for every $\alpha \in \Gamma$ and $v \in \mathcal{V}_\alpha$ we can construct, on the same probability space, the indicators $\{I''_{\beta iv}(\alpha), \beta \in \Gamma, i = 1, ..., |\Gamma_\alpha^{vs}| + 1\}$ and $\{I'_{\beta v}(\alpha), \beta \in \Gamma\}$ in such a way that*

$$\mathcal{L}(I''_{\beta iv}(\alpha), \beta \in \Gamma) = \mathcal{L}(I_\beta, \beta \in \Gamma \,\big|\, I_\alpha \mathbf{1}_{\{Z_\alpha = i\}} = 1, V_\alpha = v), \forall i \tag{1.5}$$

$$\mathcal{L}(I'_{\beta v}(\alpha), \beta \in \Gamma) = \mathcal{L}(I_\beta, \beta \in \Gamma). \tag{1.6}$$

Then, for all choices of the sets Γ_α^{vs} et Γ_α^{vw} and for all bounded functions $g : \mathbb{N} \longrightarrow \mathbb{R}$, we have

$$\begin{aligned}
\left|\mathbb{E}\left\{Wg(W) - \sum_{i=1}^{\infty} i\lambda_i g(W+i)\right\}\right| \leq & ||\Delta g|| \sum_{\alpha\in\Gamma} \left\{(\mathbb{E}I_\alpha)^2 + \mathbb{E}I_\alpha \mathbb{E}(U_\alpha + X_\alpha)\right.\\
& \left. + \mathbb{E}(I_\alpha X_\alpha) + \sum_{i=1}^{|\Gamma_\alpha^{vs}|+1} \sum_{\beta\in\Gamma_\alpha^{vw}} \mathbb{E}(I_\alpha \mathbf{1}_{\{Z_\alpha = i\}}\, \theta_{\beta,\alpha,i}(V_\alpha))\right\},
\end{aligned}$$

where $\theta_{\beta,\alpha,i}(v) = \mathbb{E}|I''_{\beta iv}(\alpha) - I'_{\beta v}(\alpha)|$, $\boldsymbol{\lambda} = \sum_{i=1}^{G+1} \lambda_i \delta_i$, $G = \max_{\alpha\in\Gamma}\{|\Gamma_\alpha^{vs}|\}$,
$\lambda_i = \frac{1}{i}\sum_{\alpha\in\Gamma} \mathbb{E}(I_\alpha \mathbf{1}_{\{Z_\alpha = i\}})$.

Using (1.2), we obtain the following estimate for the Kolmogorov distance :

$$d_K(\mathcal{L}(W), CP(\boldsymbol{\lambda})) \leq c_K(\boldsymbol{\lambda}) \sum_{\alpha\in\Gamma} \{(\mathbb{E}I_\alpha)^2 + \mathbb{E}I_\alpha \mathbb{E}(U_\alpha + X_\alpha) + \mathbb{E}(I_\alpha X_\alpha) + \sum_{i=1}^{|\Gamma_\alpha^{vs}|+1} \sum_{\beta\in\Gamma_\alpha^{vw}} \mathbb{E}(I_\alpha \mathbf{1}_{\{Z_\alpha=i\}} \theta_{\beta,\alpha,i}(V_\alpha))\}.$$

The choice of the parameters λ_i appearing in Theorem 1.1 is called the *canonical choice*.

Remark. *We have*

$$\sum_{i=1}^{G+1} i\lambda_i = \mathbb{E}(W).$$

Indeed,

$$\sum_{i=1}^{G+1} i\lambda_i = \sum_{i=1}^{G+1} \sum_{\alpha\in\Gamma} \mathbb{E}(I_\alpha \mathbf{1}_{\{Z_\alpha=i\}}) = \sum_{\alpha\in\Gamma} \sum_{i=1}^{G+1} \mathbb{E}(I_\alpha \mathbf{1}_{\{Z_\alpha=i\}}) = \sum_{\alpha\in\Gamma} \mathbb{E}(I_\alpha) = \mathbb{E}(W).$$

In practice it is not always easy to compute the canonical parameters λ_i.

The next theorem (see Theorem 4.F. in [30]) allows us to make a compound Poisson approximation with a smaller number of parameters.

Theorem 1.2. *For $\boldsymbol{\lambda} = \sum_{i=1}^{G+1} \lambda_i \delta_i$, let $\hat{\boldsymbol{\lambda}} = \sum_{i=1}^{\ell} \hat{\lambda}_i \delta_i$, with $\ell < G+1$, where $\hat{\lambda}_i = \lambda_i$ for $i = 2, ..., \ell$, $\hat{\lambda}_i = 0$ for $i \geq \ell+1$ and $\hat{\lambda}_1 = \lambda_1 + \sum_{i=\ell+1}^{G+1} i\lambda_i = \mathbb{E}(W) - \sum_{i=2}^{\ell} i\lambda_i$.*

Under the same assumptions as in Theorem 1.1, for all choices of the sets Γ_α^{vs} and Γ_α^{vw} and for all bounded functions $g : \mathbb{N} \longrightarrow \mathbb{R}$, we have :

$$\left| \mathbb{E}\left\{ Wg(W) - \sum_{i=1}^{\infty} i\lambda_i g(W+i) \right\} \right| \leq ||\Delta g|| \left\{ \sum_{\alpha\in\Gamma} \{(\mathbb{E}I_\alpha)^2 + \mathbb{E}I_\alpha \mathbb{E}(U_\alpha + X_\alpha) + \mathbb{E}(I_\alpha X_\alpha) + \sum_{i=1}^{|\Gamma_\alpha^{vs}|+1} \sum_{\beta\in\Gamma_\alpha^{vw}} \mathbb{E}(I_\alpha \mathbf{1}_{\{Z_\alpha=i\}} \theta_{\beta,\alpha,i}(V_\alpha))\} + \sum_{i=\ell+1}^{G+1} i(i-1)\lambda_i \right\}.$$

Using (1.2) and Theorem 1.2, we further obtain

Theorem 1.3. *Under the assumptions of Theorem 1.2, we have*

$$d_K(\mathcal{L}(W), CP(\hat{\boldsymbol{\lambda}})) \leq c_K(\hat{\boldsymbol{\lambda}}) \left\{ \sum_{\alpha\in\Gamma} [(\mathbb{E}I_\alpha)^2 + \mathbb{E}I_\alpha \mathbb{E}(U_\alpha + X_\alpha) + \mathbb{E}(I_\alpha X_\alpha) + \sum_{i=1}^{|\Gamma_\alpha^{vs}|+1} \sum_{\beta\in\Gamma_\alpha^{vw}} \mathbb{E}(I_\alpha \mathbf{1}_{\{Z_\alpha=i\}} \theta_{\beta,\alpha,i}(V_\alpha))] + \sum_{i=\ell+1}^{G+1} i(i-1)\lambda_i \right\}.$$

In the next section we apply the Stein-Chen method for compound Poisson approximation to our W_m, using Theorem 1.3.

1.4 Compound Poisson approximation for $\mathbb{P}(W_m \geq 1)$

1.4.1 The circular case

We place ourselves in the asymptotic settings of $n \longrightarrow \infty$ and $r \longrightarrow 0$ such that $nr \longrightarrow 0$.

We define $I_k := \mathbf{1}_{A_k}, k = 1, ..., n$. We recall that

$$A_k = \{\mu_m([U_{(k)}, U_{(k)} + r]) \geq h\},$$

where

$$\mu_m = \sum_{i=1}^{n} L_i \delta_{U_{(i)}}.$$

For every $k \in \{1, ..., n\}$, we choose the dependence sets as follows :

$$\begin{aligned}
\Gamma_k^{vs} &:= \{k - h_* + 2, ..., k - 1, k + 1, ..., k + h_* - 2\},\\
\Gamma_k^{vw} &:= \{j : |j - k| \ > 2(h^* - 2)\},\\
\Gamma_k^{b} &:= \Gamma \backslash \{\{\alpha\} \cup \Gamma_\alpha^{vs} \cup \Gamma_\alpha^{vw}\} = \{j : h_* - 2 < |j - k| \ \leq 2(h^* - 2)\}.
\end{aligned}$$

Here, $G = \max_{k=1,...,n} |\Gamma_k^{vs}| = 2(h_* - 2)$.

We recall that $Z_k = I_k + \sum_{j \in \Gamma_k^{vs}} I_j$. In our case $Z_k = \sum_{j=k-h_*+2}^{k+h_*-2} I_j$.

We will explicitly construct the coupling described in Theorem 1.1.

Let us define the spacings

$$S_j := U_{(j+1)} - U_{(j)}, \ j = 1, ..., n,$$

with the circular convention modulo n.

Notation. *For a sequence $(a_j)_j$, we will denote $a_{i,k} := a_i + \cdots + a_{i+k-1}$.*

For every $k \in \{1, ..., n\}$ and $\boldsymbol{\ell} \in \Lambda$ we have

$$A_k \cap \{\boldsymbol{L} = \boldsymbol{\ell}\} = \{S_k + \cdots + S_{k+h_k(\boldsymbol{\ell})-2} \leq r\} = \{S_{k,h_k(\boldsymbol{\ell})-1} \leq r\}.$$

Let $k \in \{1, ..., n\}$ be fixed.

The indicators appearing in the expression of Z_k are those from I_{k-h_*+2} to I_{k+h_*-2}.

Consequently, if $\boldsymbol{L} = \boldsymbol{\ell}$, then the spacings appearing in the expression of Z_k are $S_{k-h_*+2}, ..., S_{k+h_*+h_{k+h_*-2}(\boldsymbol{\ell})-4}$.

Let

$$V_k := (\boldsymbol{L}, S_{k-h_*+2}, ..., S_{k+h_*+h^*-4}).$$

Note that V_k contains all the spacings which may appear in the expression of Z_k, for different values of $\boldsymbol{\ell}$.

For every $v = (\boldsymbol{\ell}, z_1, ..., z_{2h_*+h^*-5})$, with $\boldsymbol{\ell} \in \Lambda, \ z_1, ..., z_{2h_*+h^*-5} > 0$ and $z_1 + \cdots + z_{2h_*+h^*-5} < 1$, we will construct on the same probability space the indicators $\{I''_{jiv}(k), j = 1, ..., n\}$ and $\{I'_j(k), j = 1, ..., n\}$ (not depending on v) verifying the relations (1.5) and (1.6) in Theorem 1.1.

We note that the event $\{I_k \mathbf{1}_{\{Z_k=i\}} = 1\}$ is V_k-measurable and thus, for having the condition (1.5) fulfilled, it suffices to construct the indicators $\{I''_{jv}(k), j = 1, ..., n\}$ (not depending on i), s.t.

$$\begin{aligned}&\mathcal{L}(I''_{jv}(k), j = 1, ..., n)\\ &= \mathcal{L}(I_j, j = 1, ..., n \,|\boldsymbol{L} = \boldsymbol{\ell}, S_{k-h_*+2} = z_1, ..., S_{k+h_*+h^*-4} = z_{2h_*+h^*-5}).\end{aligned}$$

Let $U'_1, ..., U'_n$ be independent on $\boldsymbol{L}$ and such that

$$\mathcal{L}(U'_1, ..., U'_n) = \mathcal{L}(U_{(1)}, ..., U_{(n)}).$$

Define the corresponding spacings $S'_j = U'_{j+1} - U'_j, \forall j = 1, ..., n$ (with the circular convention $U'_{n+1} = U'_1$). We then have

$$\mathcal{L}(S'_1, ..., S'_n) = \mathcal{L}(S_1, ..., S_n).$$

For $v = (\boldsymbol{\ell}, z_1, ..., z_{2h_*+h^*-5})$ with $\boldsymbol{\ell} \in \Lambda,\ z_1, ..., z_{2h_*+h^*-5} > 0$ and $z_1 + \cdots + z_{2h_*+h^*-5} < 1$, we let

$$\begin{aligned}(1.7)\qquad S''_j &= \frac{1 - \sum_{i=1}^{2h_*+h^*-5} z_i}{1 - \sum_{i=k-h_*+2}^{k+h_*+h^*-4} S'_i} S'_j,\ j \in \{1, ..., n\}\backslash\{k - h_* + 2, ..., k + h_* + h^* - 4\},\\ S''_{k-h_*+2} &= z_1, ..., S''_{k+h_*+h^*-4} = z_{2h_*+h^*-5}.\end{aligned}$$

Note that

$$\mathcal{L}(S''_1, ..., S''_n) = \mathcal{L}(S_1, ..., S_n \,|S_{k-h_*+2} = z_1, ..., S_{k+h_*+h^*-4} = z_{2h_*+h^*-5}).$$

Indeed, the $S''_j, j \in \{1, ..., n\}\backslash\{k-h_*+2, ..., k+h_*+h^*-4\}$ are distributed as the spacings generated by $n-(2h_*+h^*-5)-1$ points i.i.d. uniformly distributed on $[0, 1 - \sum_{i=1}^{2h_*+h^*-5} z_i]$ (see Lemma 1.4(d) below).

Let also

$$\mu'_m := \sum_{i=1}^{n} L_i \delta_{U'_i}.$$

For every $j \in \{1, ..., n\}$ we construct the indicators needed in Theorem 1.3 as follows :

$$\begin{aligned}I'_j(k) &:= \mathbf{1}_{\{\mu'_m([U'_j, U'_j+r]) \geq h\}},\\ I''_{jv}(k) &:= \mathbf{1}_{\{S''_j + \cdots + S''_{j+h_j(\boldsymbol{\ell})-2} \leq r\}}.\end{aligned}$$

We recall that $\boldsymbol{L}$ is independent of $(S_1, ..., S_n)$.

It is easy to see that the indicators defined above verify the conditions (1.5) and (1.6) and hence we can apply Theorem 1.3.

It remains to compute all the quantities appearing therein.

The canonical choice for the parameters of the compound Poisson distribution is :

$$\boldsymbol{\lambda} = \sum_{i=1}^{2h_*-3} \lambda_i \delta_i,\ \lambda_i = \frac{1}{i}\sum_{k=1}^{n} \mathbb{E}(I_k \mathbf{1}_{\{Z_k=i\}}).$$

In our approximation we will use only half of the parameters, by truncating at $\ell = h_*-1$. Instead of $\boldsymbol{\lambda}$ we will use

$$\hat{\boldsymbol{\lambda}} := \sum_{i=1}^{h_*-1} \hat{\lambda}_i \delta_i,$$

where $\hat{\lambda}_i = \lambda_i$ for $i = 2, ..., h_* - 1$ and

$$\hat{\lambda}_1 = \lambda_1 + \sum_{i=h_*}^{2h_*-3} i\lambda_i = \mathbb{E}(W_m) - \sum_{i=2}^{h_*-1} \lambda_i.$$

We will approximate the desired probability $\mathbb{P}(W_m \geq 1)$ by

$$p = 1 - \exp\{-\sum_{i=1}^{h_*-1} \hat{\lambda}_i\}.$$

Remark. *As the indicators $\{I''_{jv}(k), j = 1, ..., n\}$ do not depend on i, also the term $\theta_{j,k}(v) = \mathbb{E}\left|I''_{jv}(\alpha) - I'_{jv}(\alpha)\right|$ appearing in Theorem 1.3 does not depend on i and thus we obtain :*

$$d_K(\mathcal{L}(W_m), CP(\hat{\boldsymbol{\lambda}})) \leq c_K(\hat{\boldsymbol{\lambda}}) \left\{\sum_{k=1}^{n}\{(\mathbb{E}I_k)^2 + \mathbb{E}I_k\mathbb{E}(U_k + X_k) + \mathbb{E}(I_k X_k) + \sum_{j\in\Gamma_k^{vw}} \mathbb{E}(I_k\theta_{j,k}(V_k))\} + \sum_{i=h_*}^{2h_*-3} i(i-1)\lambda_i\right\},$$

where

$$U_k = \sum_{j\in\Gamma_k^{vs}} I_j = \sum_{j=k-h_*+2}^{k+h_*-2} I_j - I_k,\ Z_k = \sum_{j=k-h_*+2}^{k+h_*-2} I_j,$$

$$X_k = \sum_{j\in\Gamma_k^{b}} I_j = \sum_{j=k-2h^*+4}^{k-h_*+1} I_j + \sum_{j=k+h_*-1}^{k+2h^*-1} I_j.$$

In the next lemma we recall some classic results about order statistics on the unit circle. These results follow easily from the analogous ones about order statistics on $[0, 1]$ (see Lemma 1.11 from Appendix 1.A), by noting that the distribution of the spacings generated by n points i.i.d. uniformly on the circle of length 1 is the same as the distribution of $n-1$ points i.i.d. uniformly on $[0, 1]$.

Lemma 1.4. *Let $U_{(1)} < \cdots < U_{(n)}$ be the order statistics of n i.i.d. r.v.'s uniformly distributed on the circle of length 1. Define the spacings $S_j := U_{(j+1)} - U_{(j)}, j = 1, ..., n$, with the circular convention $U_{(n+1)} = U_{(1)}$.*

Then the r.v.'s $S_1, ..., S_n$ are exchangeable and for every $\ell \leq n-1$ we have the following properties :

(a) *The density of* $(S_1, ..., S_\ell)$ *is*

$$f(z_1, ..., z_\ell) = \frac{(n-1)!}{(n-1-\ell)!}(1-(z_1+\cdots+z_\ell))^{n-1-\ell},$$
$$0 \leq z_1+\cdots+z_\ell \leq 1.$$

(b) $S_{1,\ell} = S_1+\cdots+S_\ell$ *has the* $Beta(\ell, n-\ell)$ *distribution and hence the density*

$$f_{S_{1,\ell}}(u) = \frac{(n-1)!}{(\ell-1)!(n-1-\ell)!}u^{\ell-1}(1-u)^{n-1-\ell},\ 0 \leq u \leq 1.$$

(c) *Conditional on* $S_{1,\ell-1}$ *the spacings* $S_\ell, ..., S_n$ *are distributed as the spacings generated by* $n-\ell$ *i.i.d. r.v.'s uniformly distributed on* $[S_{1,\ell-1}, 1]$.

(d) *We have*

$$\mathcal{L}\left(\frac{S_\ell}{1-S_{1,\ell-1}}, ..., \frac{S_{n-1}}{1-S_{1,\ell-1}}\right) = \mathcal{L}(S'_1, ..., S'_{n-\ell}),$$

where $S'_j, j = 1, ..., n-\ell+1$ *are the spacings generated by* $n-\ell$ *i.i.d. r.v.'s uniformly distributed on* $[0, 1]$.

We will need also the following lemma.

Lemma 1.5. *For fixed* k, *assume that* $n \longrightarrow \infty, r \longrightarrow 0$ *s.t.* $nr \longrightarrow 0$. *Then, uniformly with respect to* $0 < nr < 1$, *we have*

$$\mathbb{P}(S_{1,k} \leq r) = \frac{(nr)^k}{k!}(1+\mathcal{O}(\frac{1}{n})+\mathcal{O}(nr))$$

and for fixed i *and* j,

$$\text{if } i<k: \mathbb{P}(S_{1,i} \leq r, S_{k,j} \leq r) = \frac{(nr)^{i+j}}{i!j!}(1+\mathcal{O}(\frac{1}{n})+\mathcal{O}(nr)),$$
$$\text{if } i \geq k: \mathbb{P}(S_{1,i} \leq r, S_{k,j} \leq r) = \frac{(2k-i+j-2)!}{(k+j-1)!(k-1)!(k-i+j-1)!}(nr)^{k+j-1}$$
$$\times (1+\mathcal{O}(\frac{1}{n})+\mathcal{O}(nr)).$$

Proof. We have

$$\begin{aligned}\mathbb{P}(S_{1,k} \leq r) &= \int_0^r \frac{(n-1)!}{(k-1)!(n-k-1)!}u^{k-1}(1-u)^{n-k-1}du \\ &= \frac{(n-1)!}{(k-1)!(n-k-1)!n^k}\int_0^{nr} x^{k-1}\left(1-\frac{x}{n}\right)^{n-k-1}dx \\ &= \int_0^{nr}\frac{x^{k-1}}{(k-1)!}(1+\mathcal{O}(\frac{1}{n})+\mathcal{O}(nr))dx \\ &= \frac{(nr)^k}{k!}(1+\mathcal{O}(\frac{1}{n})+\mathcal{O}(nr)).\end{aligned}$$

If $i < k$, by Lemma 1.4(c), the conditional distribution of $S_{i+1,j}$ given $S_{1,i}$ is the same as the distribution of a sum of j spacings determined by $n-i-1$ points uniformly distributed on $[S_{1,i}, 1]$. Hence, by conditioning on $S_{1,i}$, we obtain :

$$\begin{aligned}
\mathbb{P}(S_{1,i} \leq r, S_{k,j} \leq r) &= \mathbb{P}(S_{1,i} \leq r, S_{i+1,j} \leq r) \text{ (by exchangeability)} \\
&= \int_0^r \frac{(n-1)!}{(i-1)!(n-i-1)!} u^{i-1}(1-u)^{n-i-1} \\
&\quad \int_0^r \frac{(n-i-1)!}{(j-1)!(n-i-j-1)!} v^{j-1} \frac{(1-u-v)^{n-i-j-1}}{(1-u)^{n-i-1}} dvdu \\
&= \int_0^{nr} \frac{x^{i-1}}{(i-1)!} \int_0^{nr} \frac{y^{j-1}}{(j-1)!} \left(1 - \frac{x+y}{n}\right)^{n-i-j-1} \\
&\quad \times (1 + \mathcal{O}(\frac{1}{n}))dydx \\
&= \int_0^{nr} \frac{x^{i-1}}{(i-1)!} \int_0^{nr} \frac{y^{j-1}}{(j-1)!} (1 + \mathcal{O}(\frac{1}{n}) + \mathcal{O}(nr))dydx \\
&= \frac{(nr)^{i+j}}{i!j!}(1 + \mathcal{O}(\frac{1}{n}) + \mathcal{O}(nr)).
\end{aligned}$$

If $i \geq k$, we first condition on $S_{k,i-k+1}$ and then on $S_{1,k-1}$. By Lemma 1.4(c) we obtain :

$$\begin{aligned}
&\mathbb{P}(S_{1,i} \leq r, S_{k,j} \leq r) \\
&= \mathbb{P}(S_{1,k-1} + S_{k,i-k+1} \leq r, S_{k,i-k+1} + S_{i+1,k-i+j-1} \leq r) \\
&= \int_0^r \frac{(n-1)!}{(i-k)!(n-i+k-2)!} u^{i-k}(1-u)^{n-i+k-2} \\
&\quad \int_0^{r-u} \frac{(n-i+k-2)!}{(k-2)!(n-i-1)!} v^{k-2} \frac{(1-u-v)^{n-i-1}}{(1-u)^{n-i+k-2}} \\
&\quad \int_0^{r-u} \frac{(n-i-1)!}{(k-i+j-2)!(n-k-j)!} w^{k-i+j-2} \frac{(1-u-v-w)^{n-k-j}}{(1-u-v)^{n-i-1}} dwdvdu \\
&= \int_0^{nr} \frac{x^{i-k}}{(i-k)!} \int_0^{nr-x} \frac{y^{k-2}}{(k-2)!} \int_0^{nr-x} \frac{z^{k-i+j-2}}{(k-i+j-2)!} \\
&\quad \times (1 + \mathcal{O}(\frac{1}{n}) + \mathcal{O}(nr))dzdydx \\
&= \frac{1}{(i-k)!(k-1)!(k-i+j-1)!} \int_0^{nr} x^{i-k}(nr-x)^{2k-i+j-2} \\
&\quad \times (1 + \mathcal{O}(\frac{1}{n}) + \mathcal{O}(nr))dx \\
&= \frac{(2k-i+j-2)!}{(k+j-1)!(k-1)!(k-i+j-1)!}(nr)^{k+j-1}(1 + \mathcal{O}(\frac{1}{n}) + \mathcal{O}(nr)).
\end{aligned}$$

For the passage to the last line we have used the fact that

$$\int_0^1 x^k(1-x)^j dx = \frac{k!j!}{(k+j+1)!}.$$ □

For every $k = 1, ..., n$:

$$\mathbb{E}(I_k) = \frac{1}{|\Lambda|} \sum_{\boldsymbol{\ell} \in \Lambda} \mathbb{P}(A_k | \boldsymbol{L} = \boldsymbol{\ell}) = \frac{1}{|\Lambda|} \sum_{\boldsymbol{\ell} \in \Lambda} \mathbb{P}(S_{k, h_k(\boldsymbol{\ell})-1} \leq r)$$

and for every $\boldsymbol{\ell} \in \Lambda$, using Lemma 1.5 and the exchangeability of the spacings, we have :

$$\mathbb{P}(S_{k, h_k(\boldsymbol{\ell})-1} \leq r) = \mathbb{P}(S_{1, h_k(\boldsymbol{\ell})-1} \leq r) = \frac{(nr)^{h_k(\boldsymbol{\ell})-1}}{(h_k(\boldsymbol{\ell})-1)!}(1 + \mathcal{O}(\frac{1}{n}) + \mathcal{O}(nr)).$$

We have

$$\begin{aligned} \mathbb{E}(I_k) &= \frac{1}{|\Lambda|} \sum_{\boldsymbol{\ell} \in \Lambda} \frac{(nr)^{h_k(\boldsymbol{\ell})-1}}{(h_k(\boldsymbol{\ell})-1)!}(1 + \mathcal{O}(\frac{1}{n}) + \mathcal{O}(nr)) \\ &= \frac{1}{|\Lambda|} \sum_{\boldsymbol{\ell} \in \Lambda} \frac{(nr)^{h_1(\boldsymbol{\ell})-1}}{(h_1(\boldsymbol{\ell})-1)!}(1 + \mathcal{O}(\frac{1}{n}) + \mathcal{O}(nr)) \end{aligned}$$

and hence the upper bound :

$$\mathbb{E}(I_k) \leq \frac{(nr)^{h_*-1}}{(h_*-1)!}(1 + \mathcal{O}(\frac{1}{n}) + \mathcal{O}(nr)), \tag{1.8}$$

for $0 < nr < 1$.

We make the following assumption on the data.

Assumption 1. *We assume that we have $n_{\min} \asymp n$, more precisely*

$$n_{\min} = \alpha n(1 + \mathcal{O}(\frac{1}{n})),$$

with $\alpha \leq 1$ fixed.

From Assumption 1 we obtain

$$\begin{aligned} |\{\boldsymbol{\ell} \in \Lambda : h_1(\boldsymbol{\ell}) = h_*\}| &\geq |\{\boldsymbol{\ell} \in \Lambda : \ell_1 = \cdots = \ell_{h_*} = \phi_1'| \\ &= \frac{(n-h_*)!}{(n_{\min}-h_*)!(g_2\phi_2')! \cdots (g_J\phi_J')!} \\ &= |\Lambda| \frac{(n_{\min}-h_*+1) \cdots n_{\min}}{(n-h_*+1) \cdots n} \\ &= \alpha^{h_*}|\Lambda|(1 + \mathcal{O}(\frac{1}{n})) \end{aligned}$$

and hence

$$|\{\boldsymbol{\ell} \in \Lambda : h_1(\boldsymbol{\ell}) = h_*\}| \asymp |\Lambda|. \tag{1.9}$$

This implies that

$$\mathbb{E}(I_k) = \frac{|\{\boldsymbol{\ell} : h_1(\boldsymbol{\ell}) = h_*\}|}{|\Lambda|} \frac{(nr)^{h_*-1}}{(h_*-1)!}(1 + \mathcal{O}(\frac{1}{n}) + \mathcal{O}(nr))$$

and thus

$$\alpha^{h_*}\frac{(nr)^{h_*-1}}{(h_*-1)!}(1+\mathcal{O}(\frac{1}{n})+\mathcal{O}(nr)) \leq \mathbb{E}(I_k) \leq \frac{(nr)^{h_*-1}}{(h_*-1)!}(1+\mathcal{O}(\frac{1}{n})+\mathcal{O}(nr)).$$

It follows that

$$\mathbb{E}(I_k) \asymp \frac{(nr)^{h_*-1}}{(h_*-1)!}$$

and

$$\mathbb{E}(W_m) \asymp \frac{n(nr)^{h_*-1}}{(h_*-1)!}. \tag{1.10}$$

Remark. *If $\phi_1'=1$ and h is an integer, then $h_*=h$ and we have*

$$|\{\boldsymbol{\ell} : h_1(\boldsymbol{\ell})=h_*\}| \ = \alpha^h|\Lambda|(1+\mathcal{O}(\frac{1}{n}))$$

and

$$\mathbb{E}(I_k) = \alpha^h\frac{(nr)^{h-1}}{(h-1)!}(1+\mathcal{O}(\frac{1}{n})+\mathcal{O}(nr)),$$
$$\mathbb{E}(W_m) = \alpha^h\frac{n(nr)^{h-1}}{(h-1)!}(1+\mathcal{O}(\frac{1}{n})+\mathcal{O}(nr)).$$

In what follows, we will consider the intersections two by two of the events A_k.
Let $k<j$. We have

$$\mathbb{E}(I_kI_j) = \frac{1}{|\Lambda|}\sum_{\boldsymbol{\ell}\in\Lambda}\mathbb{P}(S_{k,h_k(\boldsymbol{\ell})-1}\leq r, S_{j,h_j(\boldsymbol{\ell})-1}\leq r),$$

where for each $\boldsymbol{\ell}\in\Lambda$, using Lemma 1.5, we have

$$\begin{aligned}
&\mathbb{P}(S_{k,h_k(\boldsymbol{\ell})-1}\leq r, S_{j,h_j(\boldsymbol{\ell})-1}\leq r) = \mathbb{P}(S_{1,h_k(\boldsymbol{\ell})-1}\leq r, S_{j-k+1,h_j(\boldsymbol{\ell})-1}\leq r)\\
&= \frac{(2(j-k)+h_j(\boldsymbol{\ell})-h_k(\boldsymbol{\ell}))!}{(j-k)!(j-k+h_j(\boldsymbol{\ell})-h_k(\boldsymbol{\ell}))!(j-k+h_j(\boldsymbol{\ell})-1)!}(nr)^{j-k+h_j(\boldsymbol{\ell})-1}\\
&\quad\times(1+\mathcal{O}(\frac{1}{n})+\mathcal{O}(nr)), \qquad (1.11)\\
&\quad\text{if } k<j\leq k+h_k(\boldsymbol{\ell})-2 \text{ (the two clusters intersect)}\\
&= \frac{1}{(h_k(\boldsymbol{\ell})-1)!(h_j(\boldsymbol{\ell})-1)!}(nr)^{h_k(\boldsymbol{\ell})+h_j(\boldsymbol{\ell})-2}(1+\mathcal{O}(\frac{1}{n})+\mathcal{O}(nr)), \qquad (1.12)\\
&\quad\text{if } j>k+h_k(\boldsymbol{\ell})-2 \text{ (the two clusters do not intersect).}
\end{aligned}$$

Note that all $A_k, k=1,...,n$ have the same probability and $\mathbb{E}(I_kI_j)$ depends only on the difference $j-k$.

Remark. *From Assumption 1 we can obtain, in a similar manner to (1.9), that*

$$|\{\boldsymbol{\ell}\in\Lambda : h_k(\boldsymbol{\ell})=h_*, h_j(\boldsymbol{\ell})=h_*\}| \ \asymp |\Lambda|. \tag{1.13}$$

Next we will estimate the error terms appearing in Theorem 1.3.

We will use the following general result on the exponential distribution, which we will prove in Appendix 1.A.

Lemma 1.6. *Let $X_1, ..., X_n$ be i.i.d. r.v.'s with distribution* Exp(1) *and let $i, k \geq 1$ s.t. $i + k - 1 \leq n$. Then, uniformly in $\alpha \geq 1, \beta < 1, \alpha\beta > 2(n-k-1)$, we have the following inequality :*

$$\mathbb{P}(X_{1,n} > \alpha\beta,\ X_{i,k} < \beta) \leq 2\frac{\beta^k}{k!}\frac{(\alpha\beta)^{n-k-1}}{(n-k-1)!}e^{-\alpha\beta}.$$

In what follows we obtain estimates for the error terms appearing in Theorem 1.3. We have the following result.

Proposition 1.7. *Assume that $n \longrightarrow \infty, r \longrightarrow 0$ s.t. $nr \longrightarrow 0$ and $n_{\min} \asymp n$. Then, uniformly in $\frac{1}{n} \leq nr < 1$ and*

$$n > 2(2h_* + h^* - 4) \vee \exp\left\{\frac{4(h_* + h^* - 5)}{3(h_* - 1) + h^*}\right\},$$

we have the following estimates :

$$(a)\quad \sum_{k=1}^{n}(\mathbb{E}I_k)^2 \leq \frac{n(nr)^{2(h_*-1)}}{[(h_*-1)!]^2}(1 + \mathcal{O}(\frac{1}{n}) + \mathcal{O}(nr)).$$

$$(b)\quad \sum_{k=1}^{n}\mathbb{E}(I_k)\mathbb{E}(U_k + X_k) \leq 4(h^* - 2)\frac{n(nr)^{2(h_*-1)}}{[(h_*-1)!]^2}(1 + \mathcal{O}(\frac{1}{n}) + \mathcal{O}(nr)).$$

$$(c)\quad \sum_{k=1}^{n}\mathbb{E}(I_k X_k) \leq 2(2h^* - h_* - 2)\frac{n(nr)^{2(h_*-1)}}{[(h_*-1)!]^2}(1 + \mathcal{O}(\frac{1}{n}) + \mathcal{O}(nr)).$$

$$(d)\quad \sum_{k=1}^{n}\sum_{j\in\Gamma_k^{vw}}\mathbb{E}(I_k\theta_{j,k}(V_k)) \leq 2(h_*-1)\{2h_* + h^* - 5 + 2^{h_*-2}(h_* + h^* - 4)\} \times \frac{n(nr)^{2(h_*-1)}}{[(h_*-1)!]^2}(1 + \mathcal{O}(\frac{1}{n}) + \mathcal{O}(nr)).$$

$$(e)\quad \sum_{i=h_*}^{2h_*-3} i(i-1)\lambda_i \leq (h_*-2)2^{2h_*-5}\frac{n(nr)^{2(h_*-1)}}{[(h_*-1)!]^2}(1 + \mathcal{O}(\frac{1}{n}) + \mathcal{O}(nr)).$$

Proof.

Proof of (a) and (b).
From (1.8) we deduce (a) :

$$\sum_{k=1}^{n}(\mathbb{E}I_k)^2 \leq \frac{n(nr)^{2(h_*-1)}}{[(h_*-1)!]^2}(1 + \mathcal{O}(\frac{1}{n}) + \mathcal{O}(nr))$$

and (b) :

$$\sum_{k=1}^{n}\mathbb{E}(I_k)\mathbb{E}(U_k + X_k) = \sum_{k=1}^{n}\mathbb{E}(I_k)\sum_{j:1\leq|j-k|\leq 2(h^*-2)}\mathbb{E}(I_j) \leq 4(h^*-2)\frac{n(nr)^{2(h_*-1)}}{[(h_*-1)!]^2}(1 + \mathcal{O}(\frac{1}{n}) + \mathcal{O}(nr)).$$

Proof of (c).
We have

$$\sum_{k=1}^{n}\mathbb{E}(I_kX_k)=\sum_{k=1}^{n}\left\{\sum_{j=k-2h^*+4}^{k-h_*+1}\mathbb{E}(I_jI_k)+\sum_{j=k+h_*-1}^{k+2h^*-4}\mathbb{E}(I_kI_j)\right\}.$$

Consider the cases $j=k+h_*-1,...,k+2h^*-4$. We have

$$\mathbb{E}(I_kI_j)=\frac{1}{|\Lambda|}\sum_{\boldsymbol{\ell}\in\Lambda}\mathbb{P}(S_{k,h_k(\boldsymbol{\ell})-1}\leq r,S_{j,h_j(\boldsymbol{\ell})-1}\leq r).$$

We will treat separately the cases $j=k+h_*-1$ and $j=k+h_*,...,k+2h^*-4$.
For $j=k+h_*-1$, using Lemma 1.5, we have :
– if $\boldsymbol{\ell}$ is s.t. $h_j(\boldsymbol{\ell})=h_*$ and $h_k(\boldsymbol{\ell})=h_*$, then the two clusters do not intersect and from (1.12) we obtain

$$\mathbb{P}(S_{k,h_k(\boldsymbol{\ell})-1}\leq r,S_{j,h_j(\boldsymbol{\ell})-1}\leq r)=\frac{(nr)^{2(h_*-1)}}{[(h_*-1)!]^2}(1+\mathcal{O}(\frac{1}{n})+\mathcal{O}(nr));$$

– if $\boldsymbol{\ell}$ is s.t. $h_j(\boldsymbol{\ell})=h_*$ and $h_k(\boldsymbol{\ell})>h_*$, then the two clusters intersect (because $j-k=h_*-1\leq h_k(\boldsymbol{\ell})-2$) and we have, using (1.11),

$$\begin{aligned}&\mathbb{P}(S_{k,h_k(\boldsymbol{\ell})-1}\leq r,S_{j,h_j(\boldsymbol{\ell})-1}\leq r)\\&=\frac{(3h_*-h_k(\boldsymbol{\ell})-2)!}{(h_*-1)!(2h_*-1-h_k(\boldsymbol{\ell}))!(2h_*-2)!}(nr)^{2(h_*-1)}(1+\mathcal{O}(\frac{1}{n})+\mathcal{O}(nr))\\&\leq\frac{(2h_*-3)!}{(h_*-1)!(2h_*-2)!(h_*-2)!}(nr)^{2(h_*-1)}(1+\mathcal{O}(\frac{1}{n})+\mathcal{O}(nr))\\&=\frac{1}{2}\frac{(nr)^{2(h_*-1)}}{[(h_*-1)!]^2}(1+\mathcal{O}(\frac{1}{n})+\mathcal{O}(nr));\end{aligned}$$

– for every other $\boldsymbol{\ell}$ we have

$$\mathbb{P}(S_{k,h_k(\boldsymbol{\ell})-1}\leq r,S_{j,h_j(\boldsymbol{\ell})-1}\leq r)=(nr)^{2(h_*-1)}\mathcal{O}(nr).$$

It follows from (1.13) that

$$\mathbb{E}(I_kI_{k+h_*-1})\leq\frac{(nr)^{2(h_*-1)}}{[(h_*-1)!]^2}(1+\mathcal{O}(\frac{1}{n})+\mathcal{O}(nr)).$$

For $j=k+h_*,...,k+2h^*-4$:
– if $h_k(\boldsymbol{\ell})=h_j(\boldsymbol{\ell})=h_*$, then the two clusters do not intersect and we have

$$\mathbb{P}(S_{k,h_k(\boldsymbol{\ell})-1}\leq r,S_{j,h_j(\boldsymbol{\ell})-1}\leq r)=\frac{(nr)^{2(h_*-1)}}{[(h_*-1)!]^2}(1+\mathcal{O}(\frac{1}{n})+\mathcal{O}(nr));$$

– for every other $\boldsymbol{\ell}$ we have

$$\mathbb{P}(S_{k,h_k(\boldsymbol{\ell})-1}\leq r,S_{j,h_j(\boldsymbol{\ell})-1}\leq r)=(nr)^{2(h_*-1)}\mathcal{O}(nr).$$

From (1.13) we deduce

$$\mathbb{E}(I_kI_j) \leq \frac{(nr)^{2(h_*-1)}}{[(h_*-1)!]^2}(1+\mathcal{O}(\frac{1}{n})+\mathcal{O}(nr)).$$

The cases $j = k-2h_*+4, ..., k-h_*+1$ are similar and the upper bound stated in (c) follows.

Proof of (d).
Let us denote

$$D_2 := \sum_{k=1}^{n}\sum_{j\in\Gamma_k^{vw}} \mathbb{E}(I_k\theta_{j,k}(V_k)).$$

We will condition on the values of $V_k = (\boldsymbol{L}, S_{k-h_*+2}, ..., S_{k+h_*+h^*-4})$.
Given that $(\boldsymbol{L}, S_{k-h_*+2}, ..., S_{k+h_*+h^*-4}) = (\boldsymbol{\ell}, z_1, ..., z_{2h_*+h^*-5})$, we have
$I_k = \mathbf{1}_{\{z_{h_*-1,h_k(\boldsymbol{\ell})-1}\leq r\}}$ (hence deterministic) and we obtain

$$D_2 = \sum_{k=1}^{n}\sum_{j\in\Gamma_k^{vw}} \frac{1}{|\Lambda|}\sum_{\boldsymbol{\ell}\in\Lambda} d_2(k,j,\boldsymbol{\ell}),$$

where for each $k = 1, ..., n$, $j \in \Gamma_k^{vw}$ and $\boldsymbol{\ell} \in \Lambda$ we let

$$\begin{aligned}
&d_2(k,j,\boldsymbol{\ell}) := \mathbb{E}[I_k\theta_{j,k}(V_k)|\boldsymbol{L}=\boldsymbol{\ell}]\\
&= \int \mathbf{1}_{\{z_{h_*-1,h_k(\boldsymbol{\ell})-1}\leq r\}}\mathbb{P}(\mathbf{1}_{\{S''_{j,h_j(\boldsymbol{\ell})-1}\leq r\}} \neq \mathbf{1}_{\{S'_{j,h_j(\boldsymbol{\ell})-1}\leq r\}})dF(z_1,...,z_{2h_*+h^*-5})\\
&= d_{21}(k,j,\boldsymbol{\ell}) + d_{22}(k,j,\boldsymbol{\ell}),
\end{aligned}$$

with F being the distribution of $(S_{k-h_*+2}, ..., S_{k+h_*+h^*-4})$ and

$$\begin{aligned}
d_{21}(k,j,\boldsymbol{\ell}) &:= \int \mathbf{1}_{\{z_{h_*-1,h_k(\boldsymbol{\ell})-1}\leq r\}}\mathbb{P}(S''_{j,h_j(\boldsymbol{\ell})-1}\leq r, S'_{j,h_j(\boldsymbol{\ell})-1} > r)\\
&\qquad dF(z_1,...,z_{2h_*+h^*-5}),\\
d_{22}(k,j,\boldsymbol{\ell}) &:= \int \mathbf{1}_{\{z_{h_*-1,h_k(\boldsymbol{\ell})-1}\leq r\}}\mathbb{P}(S''_{j,h_j(\boldsymbol{\ell})-1}> r, S'_{j,h_j(\boldsymbol{\ell})-1} \leq r)\\
&\qquad dF(z_1,...,z_{2h_*+h^*-5}).
\end{aligned}$$

We will estimate each of these two terms.
We note that for $j \in \Gamma_k^{vw}$ (i.e. $j < k-2h^*+4$ or $j > k+2h^*-4$) we have

$$\{j, ..., j+h_j(\boldsymbol{\ell})-2\} \subset \{1, ..., n\}\backslash\{k-h_*+2, ..., k+h_*+h^*-4\},$$

and from (1.7) :

$$S''_{j,h_j(\boldsymbol{\ell})-1} = \frac{1-z_{1,2h_*+h^*-5}}{1-S'_{k-h_*+2,2h_*+h^*-5}}S'_{j,h_j(\boldsymbol{\ell})-1}.$$

We will further decompose

$$\begin{aligned}
d_{21}(k,j,\boldsymbol{\ell}) &= d'_{21}(k,j,\boldsymbol{\ell}) + d''_{21}(k,j,\boldsymbol{\ell}),\\
d_{22}(k,j,\boldsymbol{\ell}) &= d'_{22}(k,j,\boldsymbol{\ell}) + d''_{22}(k,j,\boldsymbol{\ell}),
\end{aligned}$$

where

$$d'_{21}(k,j,\boldsymbol{\ell}) := \int \mathbf{1}_{\{z_{h_*-1,h_k(\boldsymbol{\ell})-1}\le r\}}\mathbf{1}_{\{z_{1,2h_*+h^*-5}>ar\}}\mathbb{P}(S''_{j,h_j(\boldsymbol{\ell})-1}\le r, S'_{j,h_j(\boldsymbol{\ell})-1}>r)$$
$$dF(z_1,...,z_{2h_*+h^*-5}),$$
$$d''_{21}(k,j,\boldsymbol{\ell}) := \int \mathbf{1}_{\{z_{h_*-1,h_k(\boldsymbol{\ell})-1}\le r\}}\mathbf{1}_{\{z_{1,2h_*+h^*-5}\le ar\}}\mathbb{P}(S''_{j,h_j(\boldsymbol{\ell})-1}\le r, S'_{j,h_j(\boldsymbol{\ell})-1}>r)$$
$$dF(z_1,...,z_{2h_*+h^*-5}),$$
$$d'_{22}(k,j,\boldsymbol{\ell}) := \int \mathbf{1}_{\{z_{h_*-1,h_k(\boldsymbol{\ell})-1}\le r\}}\mathbf{1}_{\{z_{1,2h_*+h^*-5}>ar\}}\mathbb{P}(S''_{j,h_j(\boldsymbol{\ell})-1}> r, S'_{j,h_j(\boldsymbol{\ell})-1}\le r)$$
$$dF(z_1,...,z_{2h_*+h^*-5}),$$
$$d''_{22}(k,j,\boldsymbol{\ell}) := \int \mathbf{1}_{\{z_{h_*-1,h_k(\boldsymbol{\ell})-1}\le r\}}\mathbf{1}_{\{z_{1,2h_*+h^*-5}\le ar\}}\mathbb{P}(S''_{j,h_j(\boldsymbol{\ell})-1}> r, S'_{j,h_j(\boldsymbol{\ell})-1}\le r)$$
$$dF(z_1,...,z_{2h_*+h^*-5}),$$

with $a = a(n)$ to be chosen a little further.

We will simplify the notation by writing h_k instead of $h_k(\boldsymbol{\ell})$.

We will use the following result. A proof of it can be found in Roos [30].

Lemma 1.8. *For $0 \le x \le 1$ and $n \ge 2(m+1)$ we have*

$$(1-x)^{n-(m+1)} \le e^{-nx/2}.$$

We obtain :

$$
\begin{aligned}
d'_{21}(k,j,\boldsymbol{\ell}) &\leq \int \mathbf{1}_{\{z_{h_*-1,h_k-1}<r\}}\mathbf{1}_{\{z_{1,2h_*+h^*-5}>ar\}}dF(z_1,...,z_{2h_*+h^*-5}) \\
&= \int_0^r \frac{n(nu)^{h_k-2}}{(h_k-2)!}\int_{ar-u}^1 \frac{n(nv)^{2h_*+h^*-h_k-5}}{(2h_*+h^*-h_k-5)!}(1-u-v)^{n-(2h_*+h^*-4)} \\
&\quad \times (1+\mathcal{O}(\frac{1}{n}))dvdu \quad \text{(Lemma 1.4}(a),(c)\text{)} \\
&= \int_0^{nr} \frac{x^{h_k-2}}{(h_k-2)!}\int_{anr-x}^n \frac{y^{2h_*+h^*-h_k-5}}{(2h_*+h^*-h_k-5)!}\left(1-\frac{x+y}{n}\right)^{n-(2h_*+h^*-4)} \\
&\quad \times (1+\mathcal{O}(\frac{1}{n}))dydx \\
&\leq \int_0^{nr} \frac{x^{h_k-2}}{(h_k-2)!}e^{-x/2}\int_{anr-x}^n \frac{y^{2h_*+h^*-h_k-5}}{(2h_*+h^*-h_k-5)!}e^{-y/2} \\
&\quad \times (1+\mathcal{O}(\frac{1}{n}))dydx \quad \text{(Lemma 1.8)} \\
&\leq \int_0^{nr} \frac{x^{h_k-2}}{(h_k-2)!}e^{-x/2}\int_{anr-x}^\infty \frac{y^{2h_*+h^*-h_k-5}}{(2h_*+h^*-h_k-5)!}e^{-y/2}(1+\mathcal{O}(\frac{1}{n}))dydx \\
&= 2^{2h_*+h^*-5}\int_0^{nr/2} \frac{z^{h_k-2}}{(h_k-2)!}e^{-z}\int_{anr/2-z}^\infty \frac{t^{2h_*+h^*-h_k-5}}{(2h_*+h^*-h_k-5)!}e^{-t} \\
&\quad \times (1+\mathcal{O}(\frac{1}{n}))dydx \\
&\leq \frac{4(nr)^{h_k-1}}{(h_k-1)!}\frac{(anr)^{2h_*+h^*-h_k-5}}{(2h_*+h^*-h_k-5)!}e^{-anr/2}(1+\mathcal{O}(\frac{1}{n})) \quad \text{(Lemma 1.6)} \\
&\leq \frac{4(nr)^{h_k+h_*-2}}{(h_*-1)!}\frac{1}{n}\left[\frac{n}{(nr)^{h_*-1}}\frac{(anr)^{2h_*+h^*-h_k-5}}{(2h_*+h^*-h_k-5)!}e^{-anr/2}\right](1+\mathcal{O}(\frac{1}{n})) \\
&\leq \frac{1}{n}(nr)^{2(h_*-1)}\mathcal{O}(nr),
\end{aligned}
$$

if $\frac{1}{n} \leq nr$ and $(3h_*+h^*-3)\log n \leq anr \leq \sqrt{n}$, entailing $n(anr)^{2h_*+h^*-h_k-5}e^{-anr/2} \leq (nr)^{h_*}$, and if $a>1, nr<1$ and $anr > 4(2h_*+h^*-4)$ for applying Lemma 1.8 and Lemma 1.6.

The last inequality is hence valid for

$$4(2h_*+h^*-4) \vee (3h_*+h^*-3)\log n \leq anr \leq \sqrt{n} \text{ and } \frac{1}{n} \leq nr < 1.$$

Further, we have :

$$
\begin{aligned}
&d''_{21}(k,j,\boldsymbol{\ell}) \\
&= \int \mathbf{1}_{\{z_{h_*-1,h_k-1}<r\}}\mathbf{1}_{\{z_{1,2h_*+h^*-5}<ar\}} \\
&\quad \times \mathbb{P}(\frac{1-z_{1,2h_*+h^*-5}}{1-S'_{k-h_*+2,2h_*+h^*-5}}S'_{j,h_j-1}<r, S'_{j,h_j-1}>r)dF(z_1,...,z_{2h_*+h^*-5}) \\
&= \int \mathbf{1}_{\{z_{h_*-1,h_k-1}<r\}}\mathbf{1}_{\{z_{1,2h_*+h^*-5}<ar\}} \\
&\quad \times \mathbb{P}(r<S'_{j,h_j-1}<r\frac{1-S'_{k-h_*+2,2h_*+h^*-5}}{1-z_{1,2h_*+h^*-5}}, S'_{k-h_*+2,2h_*+h^*-5}<z_{1,2h_*+h^*-5}) \\
&\quad dF(z_1,...,z_{2h_*+h^*-5}) \\
&= \int_0^{nr}\frac{x^{h_k-2}}{(h_k-2)!}\int_0^{anr-x}\frac{y^{2h_*+h^*-h_k-5}}{(2h_*+h^*-h_k-5)!}\left(1-\frac{x+y}{n}\right)^{n-(2h_*+h^*-4)} \\
&\quad \int_0^{x+y}\frac{u^{2h_*+h^*-6}}{(2h_*+h^*-6)!}\int_{nr}^{nr\frac{1-\frac{u}{n}}{1-\frac{x+y}{n}}}\frac{v^{h_j-2}}{(h_j-2)!}\left(1-\frac{u+v}{n}\right)^{n-(2h_*+h^*+h_j-5)} \\
&\quad \times (1+\mathcal{O}(\frac{1}{n}))dvdudydx \\
&\leq 2^{h_j-1}\frac{1}{n}\frac{(nr)^{h_j-1}}{(h_j-2)!}\int_0^{nr}\frac{x^{h_k-2}}{(h_k-2)!}\int_0^{anr-x}\frac{y^{2h_*+h^*-h_k-5}}{(2h_*+h^*-h_k-5)!}(x+y) \\
&\quad \times \left(1-\frac{y}{n}\right)^{n-(2h_*+h^*-4)}\int_0^{x+y}\frac{u^{2h_*+h^*-6}}{(2h_*+h^*-6)!}\left(1-\frac{u}{n}\right)^{n-(2h_*+h^*+h_j-5)} \\
&\quad \times (1+\mathcal{O}(\frac{1}{n}))dudydx,
\end{aligned} \tag{1.14}
$$

if $a<\frac{1}{2r}$. The last inequality follows from

$$
\begin{aligned}
\int_{nr}^{nr\frac{1-\frac{u}{n}}{1-\frac{x+y}{n}}} v^{h_j-2}dv &\leq \left(nr\frac{1-\frac{u}{n}}{1-\frac{x+y}{n}}\right)^{h_j-2}\left(nr\frac{1-\frac{u}{n}}{1-\frac{x+y}{n}}-nr\right) \\
&\leq \frac{x+y-u}{(1-\frac{x+y}{n})^{h_j-1}}\frac{1}{n}(nr)^{h_j-1} \leq \frac{x+y}{(1-\frac{x+y}{n})^{h_j-1}}\frac{1}{n}(nr)^{h_j-1} \\
&\leq \frac{x+y}{(1-ar)^{h_j-1}}\frac{1}{n}(nr)^{h_j-1} \leq 2^{h_j-1}(x+y)\frac{1}{n}(nr)^{h_j-1}.
\end{aligned}
$$

For the integral over u and the integral over y in (1.14) we will use the following result. For every k,j with $k+j\leq n-1$, we have :

$$
\begin{aligned}
&\int_0^n \frac{w^{k-1}}{(k-1)!}\left(1-\frac{w}{n}\right)^{n-j-k-1}dw = \int_0^1 \frac{n^k}{(k-1)!}t^{k-1}(1-t)^{n-j-k-1}dt \\
&= \int_0^1 \frac{(n-j-1)!}{(k-1)!(n-j-k-1)!}t^{k-1}(1-t)^{n-j-k-1}(1+\mathcal{O}(\frac{1}{n}))dt \\
&= (1+\mathcal{O}(\frac{1}{n})).
\end{aligned} \tag{1.15}
$$

We obtain

$$\begin{aligned}
&d''_{21}(k,j,\boldsymbol{\ell})\\
&\leq 2^{h_j-1}\frac{1}{n}\frac{(nr)^{h_j-1}}{(h_j-2)!}\left(\int_0^{nr}\frac{x^{h_k-1}}{(h_k-2)!}dx+(2h_*+h^*-h_k-4)\int_0^{nr}\frac{x^{h_k-2}}{(h_k-2)!}dx\right)\\
&\quad\times(1+\mathcal{O}(\frac{1}{n}))\\
&\leq 2^{h_j-1}(2h_*+h^*-h_k-4)\frac{1}{n}\frac{(nr)^{h_j+h_k-2}}{(h_j-2)!(h_k-1)!}(1+\mathcal{O}(\frac{1}{n})+\mathcal{O}(nr)).
\end{aligned}$$

Next, we have :

$$\begin{aligned}
d'_{22}(k,j,\boldsymbol{\ell}) &= \int \mathbf{1}_{\{z_{h_*-1,h_k-1}<r\}}\mathbf{1}_{\{z_{1,2h_*+h^*-5}>ar\}}\mathbb{P}(S''_{j,h_j-1}\geq r, S'_{j,h_j-1}<r)\\
&\quad dF(z_1,...,z_{2h_*+h^*-5})\\
&\leq \int \mathbf{1}_{\{z_{h_*-1,h_k-1}<r\}}\mathbf{1}_{\{z_{1,2h_*+h^*-5}>ar\}}\mathbb{P}(S'_{j,h_j-1}<r)dF(z_1,...,z_{2h_*+h^*-5})\\
&\leq \frac{(nr)^{h_j-1}}{(h_j-1)!}\int_0^{nr}\frac{x^{h_k-2}}{(h_k-2)!}e^{-x/2}\int_{anr-x}^{n}\frac{y^{2h_*+h^*-h_k-5}}{(2h_*+h^*-h_k-5)!}e^{-y/2}\\
&\quad\times(1+\mathcal{O}(\frac{1}{n})+\mathcal{O}(nr))dydx\\
&\leq 2^{2h_*+h^*-5}\frac{(nr)^{h_j-1}}{(h_j-1)!}\int_0^{nr/2}\frac{z^{h_k-2}}{(h_k-2)!}e^{-z}\int_{anr/2-z}^{\infty}\frac{t^{2h_*+h^*-h_k-5}}{(2h_*+h^*-h_k-5)!}e^{-t}\\
&\quad\times(1+\mathcal{O}(\frac{1}{n})+\mathcal{O}(nr))dydx\\
&\leq \frac{4(nr)^{h_j+h_k-2}}{(h_j-1)!(h_k-1)!}\frac{(anr)^{2h_*+h^*-h_k-5}}{(2h_*+h^*-h_k-5)!}e^{-anr/2}(1+\mathcal{O}(\frac{1}{n})+\mathcal{O}(nr))\\
&\leq \frac{4(nr)^{2(h_*-1)}}{[(h_*-1)!]^2}\frac{1}{n}\left(\frac{n(anr)^{2h_*+h^*-h_k-5}}{(2h_*+h^*-h_k-5)!}e^{-anr/2}\right)(1+\mathcal{O}(\frac{1}{n})+\mathcal{O}(nr))\\
&\leq \frac{1}{n}(nr)^{2(h_*-1)}\mathcal{O}(nr),
\end{aligned}$$

if $\frac{1}{n}\leq nr$ and $(h_*+h^*-1)\log n\leq anr\leq\sqrt{n}$, entailing $n(anr)^{2h_*+h^*-h_k-5}e^{-\alpha nr/2}\leq nr$, and if $a>1, nr<1$ and $anr>4(2h_*+h^*-4)$ for applying Lemma 1.8 and Lemma 1.6.

The last inequality is hence valid for

$$4(2h_*+h^*-4)\vee(h_*+h^*-1)\log n\leq anr\leq\sqrt{n}\text{ and }\frac{1}{n}\leq nr<1.$$

It remains to estimate the term

$$
\begin{aligned}
&d''_{22}(k,j,\boldsymbol{\ell})\\
&=\int \mathbf{1}_{\{z_{h_*-1,h_k-1}<r\}}\mathbf{1}_{\{z_{1,2h_*+h^*-5}<ar\}}\\
&\quad\times\mathbb{P}(r\frac{1-S'_{k-h_*+2,2h_*+h^*-5}}{1-z_{1,2h_*+h^*-5}}<S'_{j,h_j-1}<r, S'_{k-h_*+2,2h_*+h^*-5}>z_{1,2h_*+h^*-5})\\
&\quad dF(z_1,...,z_{2h_*+h^*-5})\\
&=\int_0^{nr}\frac{x^{h_k-2}}{(h_k-2)!}\int_0^{anr-x}\frac{y^{2h_*+h^*-h_k-5}}{(2h_*+h^*-h_k-5)!}\left(1-\frac{x+y}{n}\right)^{n-(2h_*+h^*-4)}\\
&\quad\int_{x+y}^{n}\frac{u^{2h_*+h^*-6}}{(2h_*+h^*-6)!}\int_{nr\frac{1-\frac{u}{n}}{1-\frac{x+y}{n}}}^{nr}\frac{v^{h_j-2}}{(h_j-2)!}\left(1-\frac{u+v}{n}\right)^{n-(2h_*+h^*+h_j-5)}\\
&\quad\times(1+\mathcal{O}(\frac{1}{n}))dvdudydx\\
&\leq\int_0^{nr}\frac{x^{h_k-2}}{(h_k-2)!}\int_0^{anr-x}\frac{y^{2h_*+h^*-h_k-5}}{(2h_*+h^*-h_k-5)!}\left(1-\frac{y}{n}\right)^{n-(2h_*+h^*-4)}\\
&\quad\int_{x+y}^{n}\frac{u^{2h_*+h^*-6}}{(2h_*+h^*-6)!}\left(1-\frac{u}{n}\right)^{n-(2h_*+h^*+h_j-5)}\frac{(nr)^{h_j-2}}{(h_j-2)!}\frac{u-(x+y)}{1-\frac{x+y}{n}}\frac{nr}{n}\\
&\quad\times(1+\mathcal{O}(\frac{1}{n}))dudydx\\
&\leq\frac{1}{n}\frac{(nr)^{h_j-1}}{(h_j-2)!}\frac{2h_*+h^*-5}{1-ar}\int_0^{nr}\frac{x^{h_k-2}}{(h_k-2)!}\int_0^{anr-x}\frac{y^{2h_*+h^*-h_k-5}}{(2h_*+h^*-h_k-5)!}\\
&\quad\times\left(1-\frac{y}{n}\right)^{n-(2h_*+h^*-4)}\int_{x+y}^{n}\frac{u^{2h_*+h^*-5}}{(2h_*+h^*-5)!}\left(1-\frac{u}{n}\right)^{n-(2h_*+h^*+h_j-5)}\\
&\quad\times(1+\mathcal{O}(\frac{1}{n}))dudydx\\
&\leq\frac{2(2h_*+h^*-5)}{(h_j-2)!}\frac{(nr)^{h_j-1}}{n}\int_0^{nr}\frac{x^{h_k-2}}{(h_k-2)!}(1+\mathcal{O}(\frac{1}{n}))dx\\
&=\frac{2(2h_*+h^*-5)}{(h_j-2)!(h_k-1)!}\frac{(nr)^{h_j+h_k-2}}{n}(1+\mathcal{O}(\frac{1}{n})),
\end{aligned}
$$

if $a<\frac{1}{2r}$. For the integrals over u and over y we have used (1.15).
We note that all the above inequalities are valid if

$$4(2h_*+h^*-4)\vee(3h_*+h^*-3)\log n\leq anr\leq\sqrt{n}\text{ and }\frac{1}{n}\leq nr<1.$$

Therefore, if we take $a:=\frac{(3h_*+h^*-3)\log n}{nr}$, then, uniformly in

$$\frac{1}{n}\leq nr<1\text{ and }n>4(2h_*+h^*-4)\vee\exp\left\{\frac{4(2h_*+h^*-4)}{3h_*+h^*-3}\right\},$$

we have

$$D_2 \leq 2(h_*-1)\{2h_*+h^*-5+2^{h_*-2}(h_*+h^*-4)\}\frac{n(nr)^{2(h_*-1)}}{[(h_*-1)!]^2}$$
$$\times (1+\mathcal{O}(\frac{1}{n})+\mathcal{O}(nr)).$$

We have used again (1.13) which states that

$$|\{\boldsymbol{\ell}\in\Lambda : h_k(\boldsymbol{\ell})=h_j(\boldsymbol{\ell})=h_*\}| \asymp |\Lambda|,$$

implying that the leading terms in the inner sum are the ones corresponding to $\boldsymbol{\ell}\in\Lambda$ with $h_k(\boldsymbol{\ell})=h_j(\boldsymbol{\ell})=h_*$.

Proof of (e).
We recall that

$$i\lambda_i=\sum_{k=1}^{n}\mathbb{E}(I_k\mathbf{1}_{\{Z_k=i\}}),$$

where $Z_k=\sum_{j=k-h_*+2}^{k+h_*-2}I_j$.
For every $k=1,...,n$ we let $\mathcal{C}_{ik}$ denote the class of all the subsets of size $i-1$ of $\Gamma_k^{vs}=\{k-h_*+2,...,k-1,k+1,...,k+h_*-2\}$.
We obtain

$$i\lambda_i=\sum_{k=1}^{n}\sum_{C\in\mathcal{C}_{ik}}\mathbb{E}(I_k\prod_{t\in C}I_t\prod_{t\in\Gamma_k^{vs}\setminus C}(1-I_t))\leq\sum_{k=1}^{n}\sum_{C\in\mathcal{C}_{ik}}\mathbb{E}(I_k\prod_{t\in C}I_t)$$
$$\leq\sum_{k=1}^{n}\sum_{C\in\mathcal{C}_{ik}}\mathbb{E}(I_{\inf C}I_{\sup C}).$$

For every $k=1,...,n$ and $C\in\mathcal{C}_{ik}$ we have

$$\mathbb{E}(I_{\inf C}I_{\sup C})=\frac{1}{|\Lambda|}\sum_{\boldsymbol{\ell}\in\Lambda}\mathbb{P}(S_{\inf C,h_{\inf C}(\boldsymbol{\ell})-1}\leq r, S_{\sup C,h_{\sup C}(\boldsymbol{\ell})-1}\leq r)$$

and $h_*-1\leq i-1\leq\sup C-\inf C$.
If $h_{\inf C}(\boldsymbol{\ell})=h_{\sup C}(\boldsymbol{\ell})=h_*$, then the two clusters do not intersect and we have

$$\mathbb{P}(S_{\inf C,h_{\inf C}(\boldsymbol{\ell})-1}\leq r, S_{\sup C,h_{\sup C}(\boldsymbol{\ell})-1}\leq r)=\frac{(nr)^{2(h_*-1)}}{[(h_*-1)!]^2}(1+\mathcal{O}(\frac{1}{n})+\mathcal{O}(nr)).$$

It follows that

$$\mathbb{E}(I_{\inf C}I_{\sup C})\leq\frac{(nr)^{2(h_*-1)}}{[(h_*-1)!]^2}(1+\mathcal{O}(\frac{1}{n})+\mathcal{O}(nr))$$

and hence

$$i\lambda_i\leq\binom{2(h_*-2)}{i-1}\frac{n(nr)^{2(h_*-1)}}{[(h_*-1)!]^2}(1+\mathcal{O}(\frac{1}{n})+\mathcal{O}(nr)).$$

But

$$\sum_{i=h_*}^{2h_*-3}(i-1)\binom{2(h_*-2)}{i-1}=(2h_*-4)\sum_{j=h_*-2}^{2h_*-5}\binom{2h_*-5}{j}=2(h_*-2)\frac{2^{2h_*-5}}{2}=(h_*-2)2^{2h_*-5}$$

and we obtain

$$\sum_{i=h_*}^{2h_*-3}i(i-1)\lambda_i\leq(h_*-2)2^{2h_*-5}\frac{n(nr)^{2(h_*-1)}}{[(h_*-1)!]^2}(1+\mathcal{O}(\frac{1}{n})+\mathcal{O}(nr)).$$

This completes the proof of Proposition 1.7. □

In the following lemma we show that the chosen parameters for the approximating compound Poisson distribution verify the relation (1.3), allowing us to use the bound (1.4) of Barbour and Xia [5].

Lemma 1.9. *If $0<nr<1$ and $n_{\min}\asymp n$, then*

$$\hat{\lambda}_i\asymp n(nr)^{i+h_*-2},$$

for every $i\in\{1,...,h_*-1\}$.

If $n_{\min}\asymp n$ and $nr\leq\gamma$, where γ is a fixed constant $\gamma<1$, then

$$i\hat{\lambda}_i\geq(i+1)\hat{\lambda}_{i+1},\forall i.$$

Proof. We have

$$i\lambda_i=\sum_{k=1}^{n}\mathbb{E}(I_k\mathbf{1}_{\{Z_k=i\}}).$$

For every k,

$$\mathbb{E}(I_k\mathbf{1}_{\{Z_k=i\}})=\sum_{C\in\mathcal{C}_{ik}}\mathbb{E}(I_k\prod_{t\in C}I_t\prod_{t\in\Gamma_k^{vs}\setminus C}(1-I_t)=P_k+Q_k,$$

where

$$P_k:=\mathbb{E}(I_{k-i+1}I_{k-i+2}\cdots I_{k-1}I_k)+\mathbb{E}(I_{k-i+2}\cdots I_{k+1})+\cdots+\ \mathbb{E}(I_k\cdots I_{k+i-1})$$

and $\mathcal{C}_{ik}$ contains all the subsets of $\Gamma_k^{vs}=\{k-h_*+2,...,k-1,k+1,...,k+h_*-2\}$ with $i-1$ elements.

We have

$$i\lambda_i=\sum_{k=1}^{n}P_k+\sum_{k=1}^{n}Q_k.$$

We will show that the leading terms in $i\lambda_i$ are the ones appearing in $\sum_{k=1}^{n}P_k$, i.e. the terms which are expectations of products of i consecutive indicators.

The Q_k's contain all the remaining terms, which are of the form $\mathbb{E}(I_j\cdots I_k\cdots I_t)$ with $k-h_*+2\leq j\leq k\leq t\leq k+h_*-2$ and $t-j\geq i$.

For a term with i consecutive indicators, of the form

$$\mathbb{E}(I_j \cdots I_{j+i-1}) = \frac{1}{|\Lambda|} \sum_{\boldsymbol{\ell} \in \Lambda} \mathbb{E}(I_j \cdots I_{j+i-1} | \boldsymbol{L} = \boldsymbol{\ell}),$$

we have that for each $\boldsymbol{\ell}$ the extreme clusters intersect (because $j + i - 1 \leq j + h_j(\boldsymbol{\ell}) - 2$, as $i < h_* \leq h_j(\boldsymbol{\ell})$, $\forall j$), and hence

$$\mathbb{P}(S_j + \cdots + S_{j+i-1+h_{j+i-1}(\boldsymbol{\ell})-2} \leq r) \leq \mathbb{E}(I_j \cdots I_{j+i-1} | \boldsymbol{L} = \boldsymbol{\ell}) \leq \mathbb{E}(I_j I_{j+i-1} | \boldsymbol{L} = \boldsymbol{\ell}).$$

Both the left and the right term are of order $(nr)^{i+h_{j+i-1}-2}$ and thus

$$\mathbb{E}(I_j \cdots I_{j+i-1} | \boldsymbol{L} = \boldsymbol{\ell}) \asymp (nr)^{i+h_{j+i-1}(\boldsymbol{\ell})-2}.$$

Using again the relation (1.9), which implies that

$$|\{\boldsymbol{\ell} \in \Lambda : h_{j+i-1}(\boldsymbol{\ell}) = h_*\}| \asymp |\Lambda|,$$

we obtain $\mathbb{E}(I_j \cdots I_{j+i-1}) \asymp (nr)^{i+h_*-2}, \forall j$.

We have that $\mathbb{E}(I_{t-i+1} \cdots I_t)$ appears in the expression of exactly i terms of the sum $\sum_{k=1}^n P_k$, precisely for $k = t - i + 1, ..., t$.

Let us examine the terms appearing in the Q_k's. They are of the form $\mathbb{E}(I_j \cdots I_t)$, with $t - j \geq i$.

Let $\boldsymbol{\ell} \in \Lambda$. We have

$$\mathbb{E}(I_j \cdots I_t | \boldsymbol{L} = \boldsymbol{\ell}) \leq \mathbb{E}(I_j I_t | \boldsymbol{L} = \boldsymbol{\ell}) \asymp \begin{cases} (nr)^{h_j(\boldsymbol{\ell})+h_t(\boldsymbol{\ell})-2}, & \text{if } t - j \leq h_j(\boldsymbol{\ell}) - 2 \\ (nr)^{t-j+h_t(\boldsymbol{\ell})-1}, & \text{if } t - j > h_j(\boldsymbol{\ell}) - 2. \end{cases}$$

In both cases we have $\mathbb{E}(I_j \cdots I_t | \boldsymbol{L} = \boldsymbol{\ell}) = \mathcal{O}((nr)^{i+h_t(\boldsymbol{\ell})-1})$ and hence

$$\mathbb{E}(I_j \cdots I_t) = \mathcal{O}((nr)^{i+h_*-1}).$$

For every $t = 1, ..., n$ the number of terms of the form $\mathbb{E}(I_j \cdots I_t)$ appearing in the expression of $i\lambda_i$ does not depend on n.

Indeed, we have $h_* - 1$ values of k s.t. $k \leq t \leq k + h_* - 2$, precisely $k = t - h_* + 2, ..., t$:

– for k from $t - h_* + 2$ to $t - 1$ we have at most $\sum_{j=i}^{t-k+h_*-1} \binom{t-k+h_*-3}{j-2} \leq 2^{t-k+h_*-3}$ terms of this form in $\mathbb{E}(I_k \mathbf{1}_{\{Z_k=i\}})$;

– for $k = t$ we have at most $2^{t-k+h_*-2} = 2^{h_*-2}$.

So, in total, we have at most

$$\sum_{k=t-h_*+2}^{t-1} 2^{t-k+h_*-3} + 2^{h_*-2} = 4^{h_*-2}$$

such terms in the expression of $i\lambda_i$.

We thus obtain

$$\mathbb{E}(I_k \mathbf{1}_{\{Z_k=i\}}) \asymp (nr)^{i+h_*-2}, \forall k = 1, ..., n$$

and

$$i\lambda_i \asymp n(nr)^{i+h_*-2}, \forall i = 1, ..., h_* - 1.$$

By definition,

$$\hat{\lambda}_1 = \mathbb{E}(W_m) - \sum_{i=2}^{h_*-1} i\lambda_i$$

and from (1.10),

$$\hat{\lambda}_1 \asymp n(nr)^{h_*-1}.$$

Consequently,

$$\frac{i\hat{\lambda}_i}{(i+1)\hat{\lambda}_{i+1}} \asymp \frac{1}{nr}, \forall i = 1, ..., h_* - 2,$$

and hence $\exists \gamma < 1$ such that for $nr \leq \gamma$ we have $i\hat{\lambda}_i \geq (i+1)\hat{\lambda}_{i+1}, \forall i$. □

Gathering together all the above results, we obtain the following upper bound on the error of approximating $\mathbb{P}(W_m \geq 1)$ by

$$p = 1 - \exp\{-\sum_{i=1}^{h_*-1} \hat{\lambda}_i\}.$$

Theorem 1.10. *Suppose that $n \longrightarrow \infty, r \longrightarrow 0$ s.t. $nr \longrightarrow 0$ and $n_{\min} \asymp n$. Then, uniformly in*

$$\frac{1}{n} \leq nr < 1 \text{ et } n > 2(2h_* + h^* - 4) \vee \exp\left\{\frac{4(2h_* + h^* - 4)}{3(h_* - 1) + h^*}\right\},$$

we have :

$$|\mathbb{P}(W_m \geq 1) - p| \leq C\frac{n(nr)^{2(h_*-1)}}{[(h_* - 1)!]^2}(1 + \mathcal{O}(\frac{1}{n}) + \mathcal{O}(nr)),$$

where

$$\begin{aligned} C = {} & 4h^* - h_* - 6 + (h_* - 1)\{2h_* + h^* - 5 + 2^{h_*-2}(h_* + h^* - 4)\} \\ & + (h_* - 2)2^{2h_*-6}. \end{aligned}$$

Moreover, if $\mathbb{E}(W_m) = \pi_\infty$ is held constant when $n \longrightarrow \infty$, then

$$|\mathbb{P}(W_m \geq 1) - p| = \mathcal{O}(\frac{1}{n}).$$

1.4.2 The computation of the parameters

In practice we will use a "Markovian" approximation for computing the parameters $\hat{\lambda}_i$ of the compound Poisson distribution.

We recall that we have chosen

$$\hat{\lambda}_i = \frac{1}{i}\sum_{k=1}^{n} \mathbb{E}(I_k \mathbf{1}_{\{Z_k = i\}}),\ i = 2, ..., h_* - 1,$$

$$\hat{\lambda}_1 = \mathbb{E}(W_m) - \sum_{i=2}^{h_*-1} i\hat{\lambda}_i,$$

where

$$Z_k = \sum_{j=k-h_*+2}^{k+h_*-2} I_j.$$

We have seen that the leading terms in the expression of $\hat{\lambda}_i$ are the ones which contain products of i consecutive indicators, so we will make the following approximation :

$$\begin{aligned}\hat{\lambda}_i &\approx \frac{1}{i}\sum_{k=1}^{n}\{\mathbb{E}[(1-I_{k-i})I_{k-i+1}I_{k-i+2}\cdots I_{k-1}I_k(1-I_{k+1})] \\ &\quad + \cdots + \ \mathbb{E}[(1-I_{k-1})I_k\cdots I_{k+i-1}(1-I_{k+i})]\} \\ &= n\mathbb{E}[(1-I_1)I_2\cdots I_{i+1}(1-I_{i+2})] \\ &= n\mathbb{P}(A_1^C\cap A_2\cdots A_{i+1}\cap A_{i+2}^C).\end{aligned}$$

By using further the "Markovian"approximation :

$$\mathbb{P}(A_1^C\cap A_2\cdots A_{i+1}\cap A_{i+2}^C)\approx \mathbb{P}(A_1^C|A_2)\mathbb{P}(A_2)\mathbb{P}(A_3|A_2)\cdots\mathbb{P}(A_{i+1}|A_i)\mathbb{P}(A_{i+2}^C|A_{i+1}),$$

we obtain

$$\begin{aligned}&\hat{\lambda}_i \approx n\pi q^{i-1}(1-q)^2, \text{ for } i=2,...,h_*-1 \\ &\hat{\lambda}_1 = n\pi - \sum_{i=2}^{h_*-1} i\hat{\lambda}_i,\end{aligned}$$

where

$$\begin{aligned}\pi &= \mathbb{P}(A_1), \\ q &= \mathbb{P}(A_2|A_1).\end{aligned}$$

For computing π and q we sum over all possible labellings $\boldsymbol{\ell}$:

$$\pi = \mathbb{P}(A_1) = \frac{1}{|\Lambda|}\sum_{\boldsymbol{\ell}\in\Lambda}\mathbb{P}(A_1|\boldsymbol{L}=\boldsymbol{\ell}),$$

where

$$\mathbb{P}(A_1|\boldsymbol{L}=\boldsymbol{\ell}) = \mathbb{P}(S_{1,h_1(\boldsymbol{\ell})-1}\le r).$$

The probability $\mathbb{P}(S_{1,h_1(\boldsymbol{\ell})-1}\le r)$ is given by the $Beta(h_1(\boldsymbol{\ell})-1, n-h_1(\boldsymbol{\ell})+1)$ distribution function (see Lemma 1.4(b)).

We note that $\mathbb{P}(A_1|\boldsymbol{L}=\boldsymbol{\ell})$ depends only on $\ell_1,...,\ell_{h^*}$ (because $h_1(\boldsymbol{\ell})$ depends at most on $\ell_1,...,\ell_{h^*}$). Therefore, it suffices to sum over all different $(\ell_1,...,\ell_{h^*})$ possible. In this way the number of terms in the sum does not depend on n.

We compute $q=\mathbb{P}(A_2|A_1)=\mathbb{P}(A_1\cap A_2)/\pi$ in a similar way. We have

$$\mathbb{P}(A_1\cap A_2) = \frac{1}{|\Lambda|}\sum_{\boldsymbol{\ell}\in\Lambda}\mathbb{P}(A_1\cap A_2|\boldsymbol{L}=\boldsymbol{\ell}) = \frac{1}{|\Lambda|}\sum_{\boldsymbol{\ell}\in\Lambda}\mathbb{P}(S_{1,h_1(\boldsymbol{\ell})-1}\le r, S_{2,h_2(\boldsymbol{\ell})-1}\le r).$$

For knowing $h_1(\boldsymbol{\ell})$ and $h_2(\boldsymbol{\ell})$ it suffices to know $\ell_1,...,\ell_{h^*+1}$, so it suffices to sum over all different $(\ell_1,...,\ell_{h^*+1})$.

For calculating $\mathbb{P}(S_{1,h_1(\boldsymbol{\ell})-1}\le r)$ and $\mathbb{P}(S_{1,h_1(\boldsymbol{\ell})-1}\le r, S_{2,h_2(\boldsymbol{\ell})-1}\le r)$ we use the formulas in Lemma 1.13 from Appendix 1.A.

1.4.3 The linear case

Next we consider the case of a linear genome.

As in the circular case, we will simplify the model and see the genome B as the interval $[0,1]$ and the positions of the n orthologs as i.i.d. r.v.'s uniformly distributed on $[0,1]$.

As before, A_k will denote the event of having in the genome B a cluster of type $(h:r)$ starting with the k-th orthologous gene, which is in position $U_{(k)}$:

$$A_k = \{\mu_m([U_{(k)}, U_{(k)} + r]) \geq h\}.$$

The difference with respect to the circular case is that we have a smaller number of possible events, precisely $n - h_* + 1$, and we have also a boundary effect which consists in the fact that for $k = n - h^* + 2, ..., n - h_* + 1$ the events A_k have a smaller probability :

$$\begin{aligned}\mathbb{P}(A_1) = \cdots = \mathbb{P}(A_{n-h^*+1}) \geq \mathbb{P}(A_{n-h^*+2}) \geq \cdots \geq \mathbb{P}(A_{n-h_*+1}) \\ > \mathbb{P}(A_{n-h_*+2}) = \cdots = \mathbb{P}(A_n) = 0.\end{aligned}$$

This happens because for the events close to the boundary we have a smaller number of labellings $\boldsymbol{\ell}$ which are allowed.

The number of clusters of type $(h:r)$ in the genome B is

$$W_m = \sum_{k=1}^{n-h_*+1} 1_{A_k}.$$

Similarly to the circular case, we approximate the distribution of W_m by a compound Poisson distribution of parameter

$$\hat{\boldsymbol{\lambda}} = \sum_{i=1}^{h_*-1} \hat{\lambda}_i \delta_i,$$

where

$$\hat{\lambda}_i = \frac{1}{i} \sum_{k=1}^{n-h_*+1} \mathbb{E}(I_k \mathbf{1}_{\{Z_k = i\}}),\ i = 2, ..., h_* - 1,$$

$$\hat{\lambda}_1 = \mathbb{E}(W_m) - \sum_{i=1}^{h_*-1} i\hat{\lambda}_i$$

and

$$Z_k = \sum_{j=k-h_*+2}^{k+h_*-2} I_j.$$

As in the circular case, we approximate $\mathbb{P}(W_m \geq 1)$ by

$$p = 1 - \exp\{-\sum_{i=1}^{h_*-1} \hat{\lambda}_i\}.$$

Notice that the error bound in Theorem 1.10 is valid in this case, too.

For the computation of the parameters we will ignore the boundary effects and will use a Markovian approximation, as before :

$$\hat{\lambda}_i \approx (n - h_* + 1)\pi q^{i-1}(1-q)^2,\ i = 2, ..., h_* - 1$$
$$\hat{\lambda}_1 = (n - h_* + 1)\pi - \sum_{i=2}^{h_*-1} i\hat{\lambda}_i,$$

where

$$\pi = \mathbb{P}(A_1),$$
$$q = \mathbb{P}(A_2|A_1)$$

are computed as in the circular case, by taking the sum over all possible labellings $\boldsymbol{\ell}$ and using the formulas in Lemma 1.12 from Appendix 1.A.

Note that, based on Assumption 1 (see also (1.9) and (1.13)), the error introduced in the computation of the parameters by ignoring the boundary effects is negligible.

1.5 Numerical results

We denote by $\boldsymbol{\phi}' := (\phi'_1, ..., \phi'_J)$ the vector containing all the distinct values $\phi'_1 < \cdots < \phi'_J$ for the sizes of the multigene families in the genome B, and we denote by $\boldsymbol{g} := (g_1, ..., g_J)$ the vector containing their multiplicities.

We present two sets of numerical results for our compound Poisson approximation, see the two tables below. In Tab. 1.1 we give the results in the circular case and in Tab. 1.2 in the linear case.

We have selected values for $\boldsymbol{\phi}', \boldsymbol{g}, h, r$ which are interesting in practice, for our biological purpose of statistically testing the significance of gene clusters found by the reference region approach.

In both tables, p is our compound Poisson approximation for the probability of interest $\mathbb{P}(W_m \geq 1)$ and $\hat{p}_{MC} \pm \varepsilon$ is a Monte Carlo estimate, based on 10^6 simulations, of the 95%-confidence interval for the same probability. We have estimated ε using the Central Limit Theorem.

Notice that, although Theorem 1.10 does not apply very well for these selected values and the theoretical bound given by the theorem is poor, the numerical results are very satisfactory.

TAB. 1.1 – Results in the circular case.

$(\boldsymbol{\phi}', \boldsymbol{g}, h, r)$	$\hat{p}_{MC} \pm \varepsilon$ (10^6 simulations)	p
$(1, 100, 8, 0.01)$	0.0053 ± 0.000146	0.0053
$(1, 100, 8, 0.012)$	0.0150 ± 0.000245	0.0153
$((1, 2), (100, 10), 8, 0.01)$	0.0061 ± 0.000156	0.0060
$((1, 2), (100, 10), 8, 0.012)$	0.0174 ± 0.000264	0.0178
$((1, 2, 3), (100, 15, 5), 8, 0.01)$	0.0070 ± 0.000167	0.0071
$((1, 2, 3), (100, 15, 5), 8, 0.02)$	0.0208 ± 0.000288	0.0212
$((1, 2, 3, 4), (100, 15, 5, 3), 8, 0.01)$	0.0072 ± 0.000170	0.0073
$((1, 2, 3, 4), (100, 15, 5, 3), 8, 0.012)$	0.0219 ± 0.000296	0.0222
$((1, 2, 3, 4, 5), (100, 15, 5, 3, 2), 8, 0.01)$	0.0071 ± 0.000169	0.0075
$((1, 2, 3, 4, 5), (100, 15, 5, 3, 2), 8, 0.012)$	0.0219 ± 0.000296	0.0227

TAB. 1.2 – Results in the linear case.

$(\boldsymbol{\phi}', \boldsymbol{g}, h, r)$	$\hat{p}_{MC} \pm \varepsilon$ (10^6 simulations)	p
$(1, 100, 8, 0.01)$	0.0052 ± 0.000144	0.0052
$(1, 100, 8, 0.012)$	0.0146 ± 0.000242	0.0151
$((1, 2), (100, 10), 8, 0.01)$	0.0060 ± 0.000155	0.0060
$((1, 2), (100, 10), 8, 0.012)$	0.0173 ± 0.000263	0.0176
$((1, 2, 3), (100, 15, 5), 8, 0.01)$	0.0068 ± 0.000165	0.0070
$((1, 2, 3), (100, 15, 5), 8, 0.02)$	0.0203 ± 0.000285	0.0211
$((1, 2, 3, 4), (100, 15, 5, 3), 8, 0.01)$	0.0072 ± 0.000170	0.0073
$((1, 2, 3, 4), (100, 15, 5, 3), 8, 0.012)$	0.0216 ± 0.000294	0.0220
$((1, 2, 3, 4, 5), (100, 15, 5, 3, 2), 8, 0.01)$	0.0073 ± 0.000171	0.0074
$((1, 2, 3, 4, 5), (100, 15, 5, 3, 2), 8, 0.012)$	0.0217 ± 0.000295	0.0226

1.A Appendix

We recall here Lemma 1.6 and we give its proof.

Lemma 1.6 *Let $X_1, ..., X_n$ be i.i.d. r.v.'s with distribution* Exp(1) *and let $i, k \geq 1$ s.t. $i + k - 1 \leq n$. Then, for $\alpha \geq 1, \beta < 1$ s.t. $\alpha\beta > 2(n-k-1)$, we have :*

$$\mathbb{P}(X_{1,n} > \alpha\beta, X_{i,k} < \beta) \leq 2\frac{\beta^k}{k!}\frac{(\alpha\beta)^{n-k-1}}{(n-k-1)!}e^{-\alpha\beta}.$$

Proof. We have

$$\begin{aligned}\mathbb{P}(X_{1,n} > \alpha\beta, X_{i,k} < \beta) &= \int_0^\beta \frac{u^{k-1}}{(k-1)!}e^{-u}\int_{\alpha\beta-u}^\infty \frac{v^{n-k-1}}{(n-k-1)!}e^{-v}dvdu \\ &= \int_0^\beta \frac{u^{k-1}}{(k-1)!}e^{-u}\sum_{j=0}^{n-k-1}\frac{(\alpha\beta-u)^j}{j!}e^{-(\alpha\beta-u)}du.\end{aligned}$$

We have used the following equality :

$$\int_a^\infty x^m e^{-x}dx = m!e^{-a}\sum_{l=0}^m \frac{a^l}{l!},$$

which can be proven by induction on m.

We obtain

$$\begin{aligned}\mathbb{P}(X_{1,n} > \alpha\beta, X_{i,k} < \beta) &\leq \frac{\beta^k}{k!}\sum_{j=0}^{n-k-1}\frac{(\alpha\beta)^j}{l!}e^{-\alpha\beta} \\ &\leq \frac{\beta^k}{k!}\frac{(\alpha\beta)^{n-k-1}}{(n-k-1)!}\left\{1 + \sum_{i=1}^\infty \left(\frac{n-k-1}{\alpha\beta}\right)^i\right\}e^{-\alpha\beta} \\ &\leq 2\frac{\beta^k}{k!}\frac{(\alpha\beta)^{n-k-1}}{(n-k-1)!}e^{-\alpha\beta}.\end{aligned}$$

□

In the next lemma we state some classic results about the distribution of uniform spacings on $[0, 1]$ (see for example Pyke [32]).

Lemma 1.11. *Let $S_j = U_{(j+1)} - U_{(j)}, j = 0, ..., n$ be the spacings generated by n i.i.d. r.v.'s uniformly distributed on $[0, 1]$, with the convention $U_{(0)} = 0$ and $U_{(n+1)} = 1$. Then the r.v.'s $S_0, ..., S_n$ are exchangeable and for every $\ell \leq n$ we have the following properties :*

(a) The density of $(S_1, ..., S_\ell)$ is

$$\begin{aligned}f(z_1, ..., z_\ell) = \frac{n!}{(n-\ell)!}(1 - (z_1 + \cdots + z_\ell))^{n-\ell}, \\ 0 \leq z_1 + \cdots + z_\ell \leq 1.\end{aligned}$$

(b) $S_{1,\ell} = S_1 + \cdots + S_\ell$ has the Beta$(\ell, n-\ell+1)$ distribution, i.e. the density

$$f_{S_{1,\ell}}(u) = \frac{n!}{(\ell-1)!(n-\ell)!}u^{\ell-1}(1-u)^{n-\ell},\ 0 \leq u \leq 1.$$

(*c*) *Conditional on $U_{(\ell)}$, the spacings $S_\ell, ..., S_n$ are distributed as the spacings generated by $n-\ell$ i.i.d. r.v.'s uniformly distributed on $[U_{(\ell)}, 1]$.*

We denote by $b(i; n, p), i = 0, ..., n$ the probabilities of the $Binomial(n, p)$ distribution :

$$b(i; n, p) = \binom{n}{i} p^i (1-p)^{n-i}.$$

We have the following results.

Lemma 1.12. *If $S_1, ..., S_n$ are the spacings generated by n r.v.'s i.i.d. uniformly in $[0, 1]$, then*

(*a*) *If $0 < r < 1$:*

$$\mathbb{P}(S_{1,h} \leq r) = \sum_{i=h}^{n} b(i; n, r).$$

(*b*) *If $r < \frac{1}{2}$:*

$$\begin{aligned} &\mathbb{P}(S_{1,h_1-1} \leq r, S_{2,h_2-1} \leq r) \\ &= \sum_{i=h_2}^{n} b(i; n, r) + \sum_{k=0}^{n-h_2} (-1)^k \binom{h_2 - h_1 + k + 1}{k+1} b(h_2 + k; n, r). \end{aligned}$$

Proof. For proving (*a*) it suffices to note that

$$\mathbb{P}(S_{1,h} \leq r) = \mathbb{P}(U_{(h+1)} - U_{(1)} \leq r) = \mathbb{P}(U_{(h)} \leq r),$$

which is the probability that among n points thrown at random in $[0, 1]$ at least h points fall in the interval $[0, r]$. The formula in (*a*) follows from the fact that the number of points falling in $[0, r]$ has a $Binomial(n, r)$ distribution.

Let us prove the second assertion (*b*).

If $h_2 = h_1 - 1$, this probability is just

$$\mathbb{P}(S_{1,h_1-1} \leq r) = \sum_{i=h_1-1}^{n} b(i; n, r), \tag{1.16}$$

using (*a*).

Consider the case $h_2 \geq h_1$.

We will calculate the desired probability by conditioning on the values of S_{1,h_2}.

Note that

$$\mathbb{P}(S_{1,h_1-1} \leq r, S_{2,h_2-1} \leq r | S_{1,h_2} \leq r) = 1. \tag{1.17}$$

Further, we have

$$\begin{aligned} &\mathbb{P}(S_{1,h_1-1} \leq r, S_{2,h_2-1} \leq r, S_{1,h_2} > r) \\ &= \int_r^{2r} \mathbb{P}(S_{1,h_1-1} \leq r, S_{2,h_2-1} \leq r | S_{1,h_2} = y) f(y) dy, \end{aligned} \tag{1.18}$$

where

$$f(y) = \frac{n!}{(h_2-1)!(n-h_2)!} y^{h_2-1}(1-y)^{n-h_2}$$

is the density of S_{1,h_2}, i.e. of the $Beta(h_2, n-h_2+1)$ distribution (see Lemma 1.4(b)).

Knowing that $S_{1,h_2} = y$, the spacings $S_1, ..., S_{h_2-1}$ are distributed as the spacings determined by h_2-1 points independently and uniformly distributed in $[0,y]$ (see Lemma 1.4(c)). Therefore, we can interpret $\mathbb{P}(S_{1,h_1-1} \leq r, S_{2,h_2-1} \leq r | S_{1,h_2} = y)$ as the probability, when throwing h_2-1 points at random in the interval $[0,y]$, that no points fall in the interval $(0, y-r)$ but at least h_1-1 points fall in the interval $(y-r, r)$.

Letting s denote the number of these points which fall in the interval $(y-r, r)$, we have

$$\begin{aligned}
&\mathbb{P}(S_{1,h_1-1} \leq r, S_{2,h_2-1} \leq r | S_{1,h_2} = y) \\
&= \sum_{s=h_1-1}^{h_2-1} \binom{h_2-1}{s} \left(\frac{2r-y}{y}\right)^s \left(\frac{y-r}{y}\right)^{h_2-1-s} \\
&= y^{-(h_2-1)} \sum_{s=h_1-1}^{h_2-1} \binom{h_2-1}{s} (2r-y)^s (y-r)^{h_2-1-s}
\end{aligned}$$

and, substituting into (1.18), we deduce

$$\begin{aligned}
&\mathbb{P}(S_{1,h_1-1} \leq r, S_{2,h_2-1} \leq r, S_{1,h_2} > r) \\
&= \frac{n!}{(n-h_2)!s!(h_2-1-s)!} \sum_{s=h_1-1}^{h_2-1} \int_r^{2r} (2r-y)^s (y-r)^{h_2-1-s} (1-y)^{n-h_2} dy.
\end{aligned}$$

Making the change of variable $z = 2r-y$ and then developing

$$(1+z-2r)^{n-h_2} = ((1-r)-(r-z))^{n-h_2},$$

the inner integral becomes

$$\begin{aligned}
&\int_0^r z^s (r-z)^{h_2-1-s} (1+z-2r)^{n-h_2} dz \\
&= \sum_{k=0}^{n-h_2} (-1)^k \binom{n-h_2}{k} (1-r)^{n-h_2-k} \int_0^r z^s (r-z)^{h_2-1-s+k} dz \\
&= \sum_{k=0}^{n-h_2} (-1)^k \binom{n-h_2}{k} (1-r)^{n-h_2-k} r^{h_2+k} \int_0^1 u^s (1-u)^{h_2-1-s+k} du \\
&= \sum_{k=0}^{n-h_2} (-1)^k \binom{n-h_2}{k} (1-r)^{n-h_2-k} r^{h_2+k} \frac{s!(h_2-1-s+k)!}{(h_2+k)!}.
\end{aligned}$$

We obtain

$$\begin{aligned}
&\mathbb{P}(S_{1,h_1-1} \leq r, S_{2,h_2-1} \leq r, S_{1,h_2} > r) \\
&= \sum_{k=0}^{n-h_2} (-1)^k \binom{n}{h_2+k} (1-r)^{n-h_2-k} r^{h_2+k} \sum_{s=h_1-1}^{h_2-1} \binom{h_2-1-s+k}{k} \\
&= \sum_{k=0}^{n-h_2} (-1)^k b(h_2+k; n, r) \binom{h_2-h_1+k+1}{k+1},
\end{aligned}$$

where we have used the binomial identity

$$\binom{k}{k} + \binom{k+1}{k} + \cdots + \binom{k+j}{k} = \binom{k+j+1}{k+1}. \tag{1.19}$$

Finally, using (1.17), we obtain the stated result :

$$\begin{aligned}
&\mathbb{P}(S_{1,h_1-1} \leq r, S_{2,h_2-1} \leq r) \\
&= \mathbb{P}(S_{1,h_2} \leq r) + \mathbb{P}(S_{1,h_1-1} \leq r, S_{2,h_2-1} \leq r, S_{1,h_2} > r) \\
&= \sum_{i=h_2}^{n} b(i; n, r) + \sum_{k=0}^{n-h_2} (-1)^k \binom{h_2-h_1+k+1}{k+1} b(h_2+k; n, r).
\end{aligned}$$

Note that the above formula is valid also in the case $h_2 = h_1 - 1$ (see the relation (1.16)), if we make the convention that a binomial coefficient $\binom{n}{k}$ equals 0 if $n < k$. □

The following analogous result for uniform spacings on the circle can be derived in a similar way using Lemma 1.4(a), or can be obtained directly from the result on uniform spacings on $[0,1]$ by replacing n by $n-1$ everywhere in the formulas in Lemma 1.12.

Lemma 1.13. *If $S_1, ..., S_n$ are the spacings generated by n r.v.'s i.i.d. uniformly on the circle of length 1, then*

(a) *If $0 < r < 1$:*

$$\mathbb{P}(S_{1,h} \leq r) = \sum_{i=h}^{n-1} b(i; n-1, r).$$

(b) *If $r < \frac{1}{2}$:*

$$\begin{aligned}
\mathbb{P}(S_{1,h_1-1} \leq r, S_{2,h_2-1} \leq r) = &\sum_{i=h_2}^{n-1} b(i; n-1, r) \\
&+ \sum_{k=0}^{n-1-h_2} (-1)^k \binom{h_2-h_1+k+1}{k+1} b(h_2+k; n-1, r).
\end{aligned}$$

Chapitre 2

Measures for the exceptionality of gene order in conserved genomic regions

2.1 Biological context and related work

In the literature there exist several statistical tests for detecting gene clusters which are significant from the point of view of the proximity of the orthologs (see [14, 23]).

But one might want to take into account also the order of the orthologs from these gene clusters, considering that the clusters where the order is more similar are even more biologically significant.

In Chapter 1 of this thesis we present a Compound Poisson approximation for computing the p-value of a given gene cluster found by the reference region approach. We considered there only the proximity of the orthologs in the cluster.

The goal of this work is to find "good" measures for quantifying the degree of conservation of the order of the orthologs in conserved genomic regions. Here, "good" means both biologically relevant and computationally accessible.

We present in this chapter three measures based on the transposition distance in the permutation group, together with analytic expressions for their distributions under the null hypothesis of random gene order. Our results may serve as a tool to increase the power of the statistical tests for detecting significant gene clusters.

We are interested in the case where the gene clusters are found by the "reference region" approach, which consists in starting with a fixed genomic region of a certain species A, called *the reference region*, and scanning the genome of another species B for significant orthologous clusters.

As before, a genome will be seen as an ordered sequence of genes, without separation into chromosomes.

In this work we treat only the case of no multigenic families, i.e. we suppose that the genes in the reference region in A have at most one ortholog in the genome B.

The literature on this subject is just at its beginnings. Sankoff and Haque [33] propose three adjacency disruption measures for comparing the order of the orthologs which are in common between two clusters in two genomes. They investigate in more detail the

“maximum adjacency disruption" criterion, giving analytic formulas for some values of its distribution under random gene order and also simulation results. They note the difficulty of taking into account, in a single statistical test, both the proximity of the orthologs and their order.

All the three measures that we propose here are based on the transposition distance in the permutation group. At first glance, the transposition distance does not seem to be very relevant from the biological point of view, as the mathematical transpositions do not correspond to any valid genomic event.

In the “genome rearrangements” literature, several more biologically relevant distances have been studied, which take into account one or a combination of different types of genomic events : reversals, translocations, chromosomic fissions and fusions, biological transpositions, block-interchanges – see [27] for a review.

The problem with using these distances as test statistics comes from the fact that their distributions for a random permutation are very difficult to obtain. In the literature there are very few results on this subject. Recently, Doignon and Labarre [13] have found the distribution of the number of alternating cycles in the bicolored breakpoint graph of a random (unsigned) permutation, which can be used to deduce the distribution of the genomic distances based on the breakpoint graph, as the block-interchange distance of Christie [9]. Sankoff and Haque [34] and Xu, Zheng and Sankoff [35], using a constructive approach, have obtained asymptotic estimates for the distribution of the number of cycles in the breakpoint graph of two random signed permutations (see Chapter 3, Section 3.1 for definitions).

In the present work we use the transposition distance in the permutation group, because it is very “nice” from the computational point of view.

But could it also be meaningful from the biological point of view ?

One positive answer to this question is given by Eriksen and Hultman in [15], where they describe an analogy between mathematical transpositions and genomic reversals. They show that the expected transposition distance in S_n after applying t random transpositions to the identity is a good approximation for the expected reversal distance of a genome with n genes (seen as a signed permutation) after applying t random reversals to the identity. By obtaining a closed formula for the first, they propose a method for estimating the true evolutionary distance between two genomes and show that this method compare well to the best results obtained with other methods.

The originality of the measures presented in Sections 2.4 and 2.5 comes from the fact that they are taking into account not only the order of the genes which are in common between the two clusters, but also the positions of the other orthologs. These measures are specifically adapted to the case where the gene clusters are found by the reference region approach.

We hope that the results obtained in this work are interesting not only in the genomic comparison context, but also in themselves, as a modest contribution to the giant pool of results about the transposition distance in the permutation group.

2.2 Mathematical framework

Let n denote the number of genes in the reference region in A which have one and only one ortholog in B.

The null hypothesis that we will consider is

$$H_0 : \text{random gene order in the genome B,}$$

under which all the $n!$ possible orderings of the n orthologs have the same probability to occur.

We label the n orthologs in such a way that their order in the reference region is given by the identity permutation Id_n, and we let π denote the permutation representing their order in the genome B.

Suppose that we are interested in the gene order in a certain conserved genomic region $\mathcal{R}$ in the genome B, which contains h orthologs and starts with the i-th ortholog, labelled $\pi(i)$.

In what follows h and i will be fixed.

We would like to find a way to quantify the conservation of the gene order in the region $\mathcal{R}$ compared to the order of the genes in the reference region.

Under the null hypothesis of random gene order in the genome B, π is a random permutation of n elements, uniformly chosen with probability $1/n!$.

For a permutation $\pi \in S_n$, we will use the notation $\pi = [\pi(1), ..., \pi(n)]$ instead of the classical notation $\begin{pmatrix} 1 & 2 & ... & n \\ \pi(1) & \pi(2) & ... & \pi(n) \end{pmatrix}$

Notation. *For a permutation $\sigma \in S_n$ and for $j \in \{1, ..., n-h+1\}$, we denote by $\sigma_{j,h}$ the restriction of σ to the set $\{j, j+1, ..., j+h-1\}$.*

Note that $\pi_{i,h}$ contains the labels of the h orthologs from the region of interest $\mathcal{R}$.

For comparing two permutations we will use the transposition distance in the symmetric group S_n, which we denote d_t.

We recall that a *transposition* in the group S_n is a cycle of length 2. The composition to the right of a given permutation $\pi = [\pi(1), ..., \pi(n)]$ with a transposition (i, j) results in the permutation

$$\pi \circ (i, j) = [\pi(1), ..., \pi(i-1), \pi(j), \pi(i+1), ..., \pi(j-1), \pi(i), \pi(j+1), ..., \pi(n)],$$

in which the elements $\pi(i)$ and $\pi(j)$ are interchanged.

The permutation group S_n is generated by the set of all the transpositions.

For two permutations $\pi, \sigma \in S_n$, the transposition distance $d_t(\pi, \sigma)$ is the minimum number of transpositions needed to transform π into σ or conversely.

We will use the following two classical results about permutations (see [12], p.118 for the first one and [10], p. 234 for the second one).

Lemma 2.1. *If σ is a permutation of n elements, then*

$$d_t(\sigma, Id_n) = n - c(\sigma),$$

where $c(\sigma)$ denotes the number of cycles in the disjoint cycle decomposition of σ.

Lemma 2.2. *The number of permutations of n elements which have k disjoint cycles is given by the* unsigned Stirling number of the first kind $s(n,k)$ *(see Definition 2.1(ii) in Appendix 2.A).*

2.3 A first distance

A first idea, and the most simple one, is to compare only the order of the h orthologous genes in the region $\mathcal{R}$, given by $\pi_{i,h}$, with their order in the reference region, given by the restriction $Id_n|_{\{\pi(i),\dots,\pi(i+h-1)\}}$.

Note that under the null hypothesis, $\pi_{i,h}$ is a random permutation of h elements chosen among n and hence, for $0 \leq d \leq h-1$, we have

$$\mathbb{P}(d_t(\pi_{i,h}, Id_n|_{\{\pi(i),\dots,\pi(i+h-1)\}}) \leq d) = \mathbb{P}(d_t(\sigma, Id_h) \leq d),$$

where σ is a random permutation of h elements.

Notation. *For $1 \leq k \leq n$, we will denote by*

$$p(n,k) := \frac{1}{n!} \sum_{j=k}^{n} s(n,j)$$

the probability that a random permutation π in S_n has at least k cycles.

Using Lemma 2.1 and Lemma 2.2, we obtain

Proposition 2.3. *For $0 \leq d \leq h-1$, we have*

$$\mathbb{P}(d_t(\pi_{i,h}, Id_n|_{\{\pi(i),\dots,\pi(i+h-1)\}}) \leq d) = p(h, h-d).$$

Notice that the distance considered here takes into account only the relative order of the h orthologs which are common to the reference region and to the orthologous region $\mathcal{R}$, and it ignores the positions of the other orthologs.

2.4 A second distance

Because we are interested in measuring the conservation of the gene order between the reference region and the region $\mathcal{R}$, another idea is to still ignore the order of the other $n-h$ orthologs in the genome B, but to take into account their positioning in the reference region.

We will consider the following "distance" :

$$d(\pi_{i,h}, Id_n) := \min\{d_t(\sigma, Id_n) : \sigma \in S_n, \sigma_{i,h} = \pi_{i,h}\},$$

i.e. we take the minimum over all possible orderings in B of the other orthologs outside the region $\mathcal{R}$.

We have the following result.

Proposition 2.4.

$$d(\pi_{i,h}, Id_n) = h - cc(\pi_{i,h}),$$

*where $cc(\pi_{i,h})$ denotes the number of cycles of π which contain only elements belonging to $\{i, ..., i+h-1\}$. We will call these cycles the "*closed cycles of $\pi_{i,h}$*".*

Proof. Indeed, by Lemma 2.1, we have

$$d(\pi_{i,h}, Id_n) = n - \max\{c(\sigma) : \sigma \in S_n, \sigma_{i,h} = \pi_{i,h}\},$$

and it suffices to note that the maximum is attained by the unique permutation σ° verifying $\sigma^\circ_{i,h} = \pi_{i,h}$ and for which the elements from $\{1, ..., n\} \setminus \{i, ..., i+h-1\}$ are all in distinct cycles. The permutation σ° has exactly $cc(\pi_{i,h}) + n - h$ cycles and hence

$$d(\pi_{i,h}, Id_n) = n - c(\sigma^\circ) = h - cc(\pi_{i,h}). \qquad \square$$

In the next theorem we give the distribution of $d(\pi_{i,h}, Id_n)$ under the null hypothesis.

Theorem 2.5. *For $0 \leq d \leq h-1$, we have*

$$\mathbb{P}(d(\pi_{i,h}, Id_n) \leq d) = \frac{1}{\binom{n}{h}} \sum_{m=h-d}^{h} \binom{n-m-1}{n-h-1} p(m, h-d).$$

Proof. Let M denote the r.v. representing the number of elements from $\{i, ..., i+h-1\}$ which are included in closed cycles of $\pi_{i,h}$.

We will compute $\mathbb{P}(d(\pi_{i,h}, Id_n) \leq d)$ by conditioning on the values of M :

$$\mathbb{P}(d(\pi_{i,h}, Id_n) \leq d) = \sum_{m=h-d}^{h} \mathbb{P}(d(\pi_{i,h}, Id_n) \leq d | M = m)\mathbb{P}(M = m). \tag{2.1}$$

The key observation is that the conditional distribution of $cc(\pi_{i,h})$ given $M = m$ is the same as the distribution of the number of cycles in a random permutation of m elements. Hence

$$\mathbb{P}(d(\pi_{i,h}, Id_n) \leq d | M = m) = p(m, h-d). \tag{2.2}$$

It remains to find the distribution of M.

Recall that, under H_0, $\pi_{i,h}$ is a random permutation of h elements chosen among n. So, for $h - d \leq m \leq h$:

$$\mathbb{P}(M = m) = \frac{|\{\pi_{i,h} : M = m\}|}{h!\binom{n}{h}}. \tag{2.3}$$

Every $\pi_{i,h}$ determines a unique permutation σ° such that $\sigma^\circ_{i,h} = \pi_{i,h}$ and having all the elements from $\{1, ..., n\} \setminus \{i, ..., i+h-1\}$ in distinct cycles.

Hence $|\{\pi_{i,h} : M = m\}|$ equals the number of permutations in S_n having all the elements from $\{1, ..., n\} \setminus \{i, ..., i+h-1\}$ in $n-h$ distinct cycles and such that exactly m elements among $i, i+1, ..., i+h-1$ do not belong to any of those cycles.

We then obtain

$$|\{\pi_{i,h} : M = m\}| = \sum_{\substack{k_1,...,k_{n-h}\geq 0 \\ k_1+\cdots+k_{n-h}=h-m}} \binom{h}{k_1,...,k_{n-h},m} m! \prod_{j=1}^{n-h} k_j! = h!\binom{n-m-1}{n-h-1}, \tag{2.4}$$

where $k_1, ..., k_{n-h}$ count the number of elements from $\{i, i+1, ..., i+h-1\}$ which are in the cycles determined, respectively, by each of the elements in $\{1, ..., n\} \setminus \{i, ..., i+h-1\}$. The multinomial coefficient stands for the number of choices for those $k_1 + \cdots + k_{n-h}$ elements. The product of factorials counts the number of different ways of forming the $n-h$ cycles. For example, the cycle number j contains $k_j + 1$ elements and there are $k_j!$ different cyclic permutations of those elements. The $m!$ term represents the number of ways of permuting the remaining elements from $\{i, i+1, ..., i+h-1\}$, elements which will form the closed cycles of $\pi_{i,h}$.

The last equality follows from the identity :

$$\left|\left\{(k_1, ..., k_\ell) : k_j \geq 0, \forall j, \sum_{j=1}^{\ell} k_j = s\right\}\right| = \binom{s+\ell-1}{\ell-1}. \tag{2.5}$$

To prove this identity, we first make the change of variables $k'_j := k_j + 1$, $j = 1, ..., \ell$ and then we use the "bars and stars" idea to notice that

$$\left|\left\{(k'_1, ..., k'_\ell) : k'_j \geq 1, \forall j, \sum_{j=1}^{\ell} k'_j = s+\ell\right\}\right|$$

represents the number of ways of separating $s+\ell$ stars arranged on a line, into ℓ nonempty groups. This number is exactly the binomial coefficient from the right-hand side of (2.5), because one has to place $\ell - 1$ bars in $\ell - 1$ of $s + \ell - 1$ places available.

From (2.3) and (2.4) we deduce that for every $m \in \{h-d, ..., h\}$:

$$\mathbb{P}(M = m) = \frac{\binom{n-m-1}{n-h-1}}{\binom{n}{h}}. \tag{2.6}$$

By substituting (2.6) and (2.2) into (2.1), the formula in the statement of the theorem follows. □

2.5 A third distance

From the biological point of view, a disadvantage of the previous distance is the fact that it is very restrictive with respect to the position of the cluster $\mathcal{R}$ in the genome B.

For taking into account eventual genomic translocations that could have changed the position of $\mathcal{R}$ with respect to the other orthologs in B, we think that a better idea would be to use the following "distance" :

$$d^*(\pi_{i,h}, Id_n) := \min\{d_t(\sigma, Id_n) : \sigma \in S_n, \sigma_{i^*,h} = \pi_{i,h}\}, \tag{2.7}$$

where

$$i^* := \arg \max_{1 \leq j \leq n-h+1} |\{\pi_i, ..., \pi_{i+h-1}\} \cap \{j, ..., j+h-1\}| .$$

We need to make a convention for the case when we have more than one maximum points. We decide, for example, to choose i^* the smallest such value.

We denote by σ° the unique permutation which attains the minimum in (2.7), hence has all the elements from $\{1, ..., n\} \setminus \{i^*, ..., i^*+h-1\}$ in distinct cycles.

As in Proposition 2.4, we have

$$d^*(\pi_{i,h}, Id_n) = h - cc(\sigma^\circ_{i^*,h}),$$

where $cc(\sigma^\circ_{i^*,h})$ denotes the number of "closed cycles of $\sigma^\circ_{i^*,h}$", i.e. the number of those cycles of σ° which do not contain any elements from $\{1, ..., n\} \setminus \{i^*, ..., i^*+h-1\}$.

Let

$$L^* := |\{\pi_i, ..., \pi_{i+h-1}\} \cap \{i^*, ..., i^*+h-1\}| .$$

Let $0 \leq d \leq h-1$. For computing the probability $\mathbb{P}(d^*(\pi_{i,h}, Id_n) \leq d)$ we will condition on the values of L^* :

$$\mathbb{P}(d^*(\pi_{i,h}, Id_n) \leq d) = \sum_{\ell=h-d}^{h} \mathbb{P}(d^*(\pi_{i,h}, Id_n) \leq d | L^* = \ell) \mathbb{P}(L^* = \ell).$$

Note that L^* must be greater than $h-d$, because we need to have at least $h-d$ closed cycles of $\sigma^\circ_{i^*,h}$ and hence at least $h-d$ elements in common between $\{\sigma^\circ_{i^*}, ..., \sigma^\circ_{i^*+h-1}\}$ and $\{i^*, ..., i^*+h-1\}$.

Next we will compute the conditional probabilities. We will prove :

Proposition 2.6. *For $0 \leq d \leq h-1$ and $h-d \leq \ell \leq h$, we have*

$$\mathbb{P}(d^*(\pi_{i,h}, Id_n) \leq d | L^* = \ell) = \frac{1}{\binom{h}{\ell}} \sum_{m=h-d}^{\ell} \binom{h-m-1}{h-\ell-1} p(m, h-d).$$

Proof. We denote by M^* the number of elements from $\{i^*, ..., i^*+h-1\}$ which are included in closed cycles of $\sigma^\circ_{i^*,h}$.

By further conditioning on M^* we obtain :

$$\mathbb{P}(d^*(\pi_{i,h}, Id_n) \leq d | L^* = \ell) = \sum_{m=h-d}^{\ell} \mathbb{P}(M^* = m | L^* = \ell) p(m, h-d). \tag{2.8}$$

Indeed, we have

$$\begin{aligned} \mathbb{P}(d^*(\pi_{i,h}, Id_n) \leq d | L^* = \ell, M^* = m) &= \mathbb{P}(d^*(\pi_{i,h}, Id_n) \leq d | M^* = m) \\ &= p(m, h-d). \end{aligned}$$

Let $\mathcal{P} = \{\pi_i, ..., \pi_{i+h-1}\}$. Under the null hypothesis, $\mathcal{P}$ is a random combination of h elements among n.

Notice that i^* and L^* are completely determined by $\mathcal{P}$ and that M^* is completely determined given a permutation of $\mathcal{P}$.

We have

$$\mathbb{P}(M^*=m|L^*=\ell)=\sum_{\mathcal{P}:L^*=\ell}\mathbb{P}(M^*=m|\mathcal{P})\mathbb{P}(\mathcal{P}|L^*=\ell). \tag{2.9}$$

Fix $\mathcal{P}$ s.t. $L^*=\ell$. Then $\mathbb{P}(M^*=m\,|\mathcal{P})$ equals the number of permutations of $\mathcal{P}$ s.t. $M^*=m$ divided by $h!$.

Let $\mathcal{I}:=\{i^*,...,i^*+h-1\}\setminus\mathcal{P}$ and $\mathcal{J}:=\mathcal{P}\setminus\{i^*,...,i^*+h-1\}$. Both $\mathcal{I}$ and $\mathcal{J}$ have $h-\ell$ elements.

Notice that the number of permutations of $\mathcal{P}$ for which $M^*=m$ equals the number of permutations σ° in S_n verifying the following conditions :

- every element from $\mathcal{I}$ is in a cycle with exactly one element from $\mathcal{J}$;
- there are no two elements of $\mathcal{I}$ or two elements of $\mathcal{J}$ in a same cycle ;
- all the elements from $\{1,...,n\}\setminus(\mathcal{P}\cup\mathcal{I}\cup\mathcal{J})$ are fixed points ;
- there are exactly m elements from $\mathcal{P}\cap\{i^*,...,i^*+h-1\}$ which do not belong to any of the $h-\ell$ cycles determined by the $h-\ell$ pairs formed by one element from $\mathcal{I}$ and one element from $\mathcal{J}$;
- $\sigma^\circ(\mathcal{J})=\mathcal{I}$.

We obtain

$$\begin{aligned}\mathbb{P}(M^*=m|\mathcal{P})&=\frac{1}{h!}(h-\ell)!\sum_{\substack{k_1,...,k_{h-\ell}\geq 0\\ k_1+\cdots+k_{h-\ell}=\ell-m}}\binom{\ell}{k_1,...,k_{h-\ell},m}m!\prod_{j=1}^{h-\ell}k_j!\\&=\frac{\binom{h-m-1}{h-\ell-1}}{\binom{h}{\ell}}\text{ (using (2.5)).}\end{aligned} \tag{2.10}$$

Indeed, we have $(h-\ell)!$ ways of pairing the elements from $\mathcal{I}$ with the elements from $\mathcal{J}$, and each of these pairings determines $h-\ell$ disjoint cycles. In the above formula, $k_1,...,k_{h-\ell}$ denote the number of elements from $\mathcal{P}\cap\{i^*,...,i^*+h-1\}$ which belong to those $h-\ell$ cycles, respectively. The multinomial coefficient stands for the choice of these elements and the product of factorials counts the number of possibilities for forming the cycles. Consider, for example, the first cycle and denote by a and b its elements from $\mathcal{I}$ and $\mathcal{J}$, respectively. This cycle contains k_1+2 elements, but we have the restriction $\sigma^\circ(b)=a$ and hence we have only $k_1!$ ways of forming it. The $m!$ term counts the number of ways of permuting the m elements which form the closed cycles of $\sigma^\circ_{i^*,h}$.

Note that the formula (2.10) is the same for all $\mathcal{P}$ satisfying $L^*=\ell$ and hence, from (2.9), we deduce

$$\mathbb{P}(M^*=m|L^*=\ell)=\frac{\binom{h-m-1}{h-\ell-1}}{\binom{h}{\ell}}.$$

Substituting into (2.8), the formula in the statement follows. □

It remains to find the distribution of

$$L^*=\max_{1\leq j\leq n-h+1}|\mathcal{P}\cap\{j,...,j+h-1\}|\,,$$

where $\mathcal{P}=\{\pi_i,...,\pi_{i+h-1}\}$ is a random combination of h elements among n.

Note that we can associate to each $\mathcal{P}$ a unique sequence $\boldsymbol{x}$ of zeros and ones, of length n, in the following manner : for every $k = 1, ..., n$, we let $x_k = 1$ if and only if $k \in \mathcal{P}$. Hence, under the null hypothesis of random gene order, $\boldsymbol{x}$ is a random sequence formed by h ones and $n - h$ zeros.

Therefore, L^* counts the maximum number of 1's within any window of length h in a random sequence formed with h 1's and $(n-h)$ 0's. This variable appears in the literature on scan statistics and it is called *the conditional discrete scan statistic*, as it is a scan statistic in the case of i.i.d. Bernoulli r.v.'s, conditional on the number of successes (1's).

Several exact formulas for the distribution of the conditional discrete scan statistic exist in the literature, for different particular cases for the parameters, and also various approximations and bounds (see [19], chapter 12).

Here we will derive an exact expression for its distribution in the most general case, by adapting the results obtained by Huntington and Naus [24] in the conditional continuous settings. We have not seen this result in the literature, although it might have already appeared. We now give a proof of it. We follow the ideas from the proof of Huntington and Naus [24] and use a result of Naus [28].

Let $X_j, j = 1, ..., n$ be i.i.d. Bernoulli random variables.

Notation. *For $1 \leq k \leq n - j + 1, j = 1, ..., n$, we denote*

$$X_{j,k} := X_j + \cdots + X_{j+k-1}.$$

Let

$$N_h := \max_{1 \leq i \leq n-h+1} X_{i,h}.$$

Notice that

$$\mathbb{P}(L^* = \ell) = \mathbb{P}(N_h = \ell | X_{1,n} = h). \tag{2.11}$$

Let $2 \leq a \leq n$. We will give the result in the general settings where we condition on having a successes (1's). We have the following result.

Proposition 2.7. *If we denote by L the integer part of $\frac{n}{h}$ and we let $b = n - Lh$, then, for $2 \leq k \leq a$:*

$$\mathbb{P}(N_h < k | X_{1,n} = a) = \frac{(b!)^{L+1}[(h-b)!]^L}{\binom{n}{a}} \sum_{Q_k} \det |d_{ij}^{(k)}| \det |g_{ij}^{(k)}|,$$

where

$$Q_k = \{(n_1, ..., n_{2L+1}) : n_i \in \mathbb{N}, \textstyle\sum_{i=1}^{2L+1} n_i = a,\ n_i + n_{i+1} < k, \forall i = 1, ..., 2L\}$$

and the determinants are of size $(L+1) \times (L+1)$ and $L \times L$ respectively, with

$$d_{ij}^{(k)} = \frac{1}{c_{ij}^{(k)}!(b - c_{ij}^{(k)})!},\ g_{ij}^{(k)} = \frac{1}{f_{ij}^{(k)}!(h - b - f_{ij}^{(k)})!},$$

where

$$c_{ij}^{(k)} = \begin{cases} -\sum_{s=2i}^{2j-2} n_s + (j-i)k, & \text{if } 1 \leq i \leq j \leq L+1 \\ \sum_{s=2j-1}^{2i-1} n_s - (i-j)k, & \text{if } 1 \leq j < i \leq L+1 \end{cases}$$

and

$$f_{ij}^{(k)} = \begin{cases} -\sum_{s=2i+1}^{2j-1} n_s + (j-i)k, \text{ if } 1 \le i \le j \le L \\ \sum_{s=2j}^{2i} n_s - (i-j)k, \text{ if } 1 \le j < i \le L. \end{cases}$$

Proof. We divide the n Bernoulli trials into $2L+1$ groups, the odd-numbered groups, $I_{2i-1}, i = 1, ..., L+1$, being of size b and the other ones, $I_{2i}, i = 1, ..., L$, of size $h-b$:

$$I_{2i-1} = \{(i-1)h+1, ..., (i-1)h+b\}, i = 1, ..., L+1$$
$$I_{2i} = \{(i-1)h+b+1, ..., ih\}, \ i = 1, ..., L.$$

For $i = 1, ..., 2L+1$, we will denote by n_i the number of 1's in the i-th group, i.e.

$$n_i = \sum_{j \in I_i} X_j.$$

Conditional on $X_{1,n} = a$, the joint distribution of the n_i's is :

$$\mathbb{P}(n_1, ..., n_{2L+1} | X_{1,n} = a) = \frac{\prod_{i=1}^{L+1} \binom{b}{n_{2i-1}} \prod_{i=1}^{L} \binom{h-b}{n_{2i}}}{\binom{n}{a}}, \text{ if } \sum_{i=1}^{2L+1} n_i = a. \tag{2.12}$$

We denote

$$S_1 = \bigcup_{i=1}^{L+1} I_{2i-1}, \ S_2 = \bigcup_{i=1}^{L} I_{2i},$$
$$m_r = \max_{i \in S_r} X_{i,h}, \ r = 1, 2.$$

Then $N_h = \max(m_1, m_2)$.

Notice that, given $\{n_i\}$, m_1 and m_2 are independent. Consequently,

$$\mathbb{P}(N_h < k | \{n_i\}) = \mathbb{P}(m_1 < k | \{n_i\}) \mathbb{P}(m_2 < k | \{n_i\}). \tag{2.13}$$

The idea is to find the conditional distributions of m_1 and m_2 given $\{n_i\}$ and then to average over the joint distribution of $\{n_i\}$.

We give here only the derivation of $\mathbb{P}(m_1 < k | \{n_i\})$, the conditional distribution of m_2 is found analogously.

For $i = 1, ..., L+1$ and $t = 1, ..., b$ we will denote

$$Y_i(t) := X_{(i-1)h+1,t}$$

the number of 1's in the first t trials of the i-th odd-numbered group I_{2i-1}. Note that $Y_i(b) = n_{2i-1}$.

The key observation is that, given $\{n_i\}$, $m_1 < k$ provided that $\{n_i\}$ is in the set Q_k defined in the statement, and further that

$$X_{(i-1)h+t+1,h} < k, \text{ for all } t = 1, ..., b-1, \ i = 1, ..., L.$$

But

$$X_{(i-1)h+t+1,h} = Y_{i+1}(t) + n_{2i} + n_{2i-1} - Y_i(t),$$

thus we can write

$$\mathbb{P}(m_1 < k|\{n_i\}) = \mathbb{P}(\bigcap_{t=1}^{b}\bigcap_{i=1}^{L}\{Y_i(t)+\alpha_i > Y_{i+1}(t)+\alpha_{i+1}\}|\{n_i\}), \tag{2.14}$$

where

$$\alpha_i - \alpha_{i+1} = k - n_{2i-1} - n_{2i} > 0,\ i = 1, ..., L$$

and hence

$$\alpha_i := (L-i+1)k - \sum_{j=2i-1}^{2L} n_j,\ i = 1, ..., L+1. \tag{2.15}$$

The probability in the right-hand side of (2.14) appears in a variant of the L-candidate ballot problem (see Naus [28]).

Using the relation (2.5) from Naus [28] we obtain

$$\mathbb{P}(m_1 < k|\{n_i\}) = \det|h_{ij}|, \tag{2.16}$$

where, for $i, j = 1, ..., L+1$:

$$h_{ij} = \frac{n_{2i-1}!(b-n_{2i-1})!}{(n_{2i-1}+\alpha_i-\alpha_j)!(b-n_{2i-1}-\alpha_i+\alpha_j)!}, \tag{2.17}$$

with the convention $h_{ij} = 0$ if any of the factorial terms is negative.

From (2.15), (2.16) and (2.17) we deduce that, for $\{n_i\}$ in Q_k, we have

$$\mathbb{P}(m_1 < k|\{n_i\}) = R\det|d_{ij}^{(k)}|, \tag{2.18}$$

where

$$R = \prod_{i=1}^{L+1} n_{2i-1}!(b-n_{2i-1})!$$

and $d_{ij}^{(k)}$ are as given in the statement.

In a similar way we can show that

$$\mathbb{P}(m_2 < k|\{n_i\}) = T\det|g_{ij}^{(k)}|, \tag{2.19}$$

where

$$T = \prod_{i=1}^{L} n_{2i}!(h-b-n_{2i})!$$

and $g_{ij}^{(k)}$ are as in the statement.

By substituting (2.18) and (2.19) into (2.13) and then averaging over the distribution of the n_i's, given in (2.12), the formula in the statement follows. □

From relation (2.11), Proposition 2.6 and Proposition 2.7 we deduce the following result about the distribution of $d^*(\pi_{i,h}, Id_n)$.

Theorem 2.8. *For $0 \leq d \leq h-1$, we have*

$$\mathbb{P}(d^*(\pi_{i,h}, Id_n) \leq d) = \frac{1}{\binom{h}{\ell}} \sum_{\ell=h-d}^{h} \mathbb{P}(L^* = \ell) \sum_{m=h-d}^{\ell} \binom{h-m-1}{h-\ell-1} p(m, \geq h-d),$$

where

$$\mathbb{P}(L^* = \ell) = \mathbb{P}(N_h < \ell+1 | X_{1,n} = h) - \mathbb{P}(N_h < \ell | X_{1,n} = h)$$

and the two conditional probabilities are given by Proposition 2.7.

Note that the values for n and h which are typical for our application are not too large (n is of the order of 100) and hence there is no problem in computing the distribution of N_h given by Proposition 2.7.

2.6 Discussion

Among the three "distances" presented in this chapter, we think that from the biological point of view the first one and the third one are the most interesting.

Our result on the distribution of the second distance may however have an interest in itself, mathematically speaking, or one may find some better applications to other problems.

While the third distance is specifically adapted to the reference region approach, the first distance could also be used in whole genome comparisons or window sampling approaches.

Based on the first idea, of comparing only the order of the orthologs which are in common between the two clusters, one could imagine replacing the transposition distance with another distance, maybe more interesting biologically. For example, we could use the block-interchange distance of Christie [9] and the results of Doignon and Labarre [13] on its distribution.

A natural continuation of this work is to try to extend these results to other distances between permutations, which may better account for the biological reality. And also, to try to obtain analogous results in the case of signed permutations (when we take into account also the orientation of the genes) and in the case of multipermutations (when we can have multiple orthologs for a given gene).

Another important question to be further considered is how to cleverly combine, in a single statistical test, the proximity of the orthologs and their order.

2.A Stirling numbers of the first kind

Definition 2.1. *We have the following equivalent definitions of the* unsigned Stirling number of the first kind $s(n,k)$:

(i) $s(n,k)$, $n \geq k \geq 0$ *satisfy the recurrence relation*

$$s(n+1,k) = ns(n,k) + s(n,k-1), \tag{2.20}$$

with boundary conditions $s(0,0) = 1$ and $s(n,0) = 0$ for $n > 0$.

(ii) For $n \geq k \geq 1$, $s(n,k)$ equals the number of permutations of n elements which have k cycles in the disjoint cycle decomposition.

(*iii*) *For* $n \geq k \geq 1$, $s(n,k)$ *is the coefficient of* x^k *in the expansion of*

$$x_{(n)} = x(x+1)\cdots(x+n-1).$$

These equivalences can be proved by showing that the Stirling numbers, as defined both in (*ii*) and in (*iii*), verify the recurrence relation (2.20). This result is checked in both cases by induction on n.

We will just show here that (*ii*) is equivalent to (*i*).

Let us consider the permutations of $n+1$ elements having k cycles. We separate them into two types, with respect to the number of cycles formed by $1, ..., n$:

- the permutations in which $1, ..., n$ already form k cycles and hence the element $(n+1)$ must belong to one of those cycles.
  For every permutation of $1, ..., n$ with k cycles, we have n different possibilities of placing the element $(n+1)$ into one of these cycles. Indeed, if the i-th cycle has length ℓ_i, we have ℓ_i possibilities to place $(n+1)$ in this cycle, so, in total, we have $\ell_1 + \cdots + \ell_k = n$ possibilities to place $(n+1)$. The permutations of this type correspond to the first term in the right-hand side of the relation (2.20).
- the permutations in which the first n elements form $k-1$ cycles, and $(n+1)$ is a fixed point. These permutations give the second term in the right-hand side of (2.20).

Chapitre 3

On the distribution of the number of cycles in the breakpoint graph of a random signed permutation

3.1 Preliminaries

We let S_n denote the permutation group of order n. For a permutation $\pi \in S_n$, we will use the notation $\pi = [\pi(1), ..., \pi(n)]$ instead the classical notation $\begin{pmatrix} 1 & 2 & ... & n \\ \pi(1) & \pi(2) & ... & \pi(n) \end{pmatrix}$.

A signed permutation of n elements is a permutation $\pi = [\pi(1), ..., \pi(n)]$ in which the elements $\pi(i), i = 1, ..., n$ have a sign, either $+$ or $-$. In other words, $\pi(i) \in \{\pm 1, ..., \pm n\}$, for $i = 1, ..., n$ and $\{|\pi(1)|, ..., |\pi(n)|\} = \{1, ..., n\}$.

Let B_n denote the set of all the signed permutations of n elements. B_n is called the *hyperoctahedral group*.

The *reversal* of the interval (i, j) in the signed permutation $\pi = [\pi(1), ..., \pi(n)]$ reverses the subsequence $\pi(i), ..., \pi(j)$ while changing their signs, hence produces the signed permutation $\pi' = [\pi(1), ..., \pi(i-1), -\pi(j), -\pi(j-1), ..., -\pi(i+1), -\pi(i), \pi(j+1), ..., \pi(n)]$.

For $\pi \in B_n$, we let $d_{rev}(\pi, Id)$ denote its reversal distance, i.e. the minimum number of reversals needed to transform π into the identity permutation $Id = [+1, ..., +n]$.

3.1.1 The breakpoint graph and the reversal distance

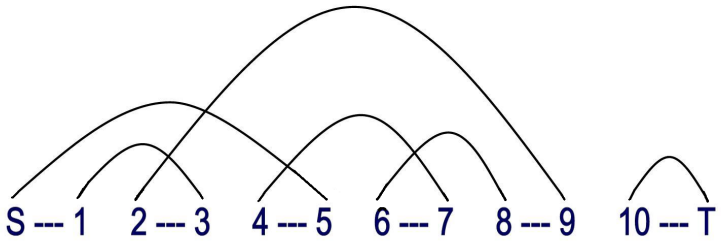
Bafna and Pevzner [3] introduced the concept of *breakpoint graph* of a permutation and noticed important links between the cycle decomposition of this graph and the reversal distance.

The breakpoint graph of a signed permutation is defined as follows.

Given a signed permutation $\pi \in B_n$, we first transform it into an unsigned permutation $\pi' \in S_{2n}$ by replacing the positive elements $+i$ by the pair $(2i - 1, 2i)$ and the negative elements $-i$ by the pair $(2i, 2i - 1)$. For instance, the signed permutation $\pi = [+3, -4, -2, +1, +5]$ is transformed into $\pi' = [5, 6, 8, 7, 4, 3, 1, 2, 9, 10]$.

We extend π' by adding two more elements, one at the beginning, denoted S (for *Start*) and one at the end, denoted T (for *Terminus*).

FIG. 3.1 – The breakpoint graph of the permutation $\pi = [+3, -4, -2, +1, +5]$.



Definition 3.1. The breakpoint graph *of the signed permutation* $\pi \in B_n$ *is the graph* $G(\pi) = (V, B \cup C)$ *which has the set of vertices* $V = \{S, 1, 2, ..., 2n, T\}$ *and the edge set partitioned into two subsets : the set* B *of solid edges, corresponding to adjacencies in the permutation* π*, and the set* C *of dashed edges, corresponding to adjacencies in the identity permutation* Id*.*

If for every element a *of the permutation* π *we denote respectively by* a_L *and* a_R *the left and right elements in the pair associated to* a *in* π'*, then we will have a solid edge between* a_R *and* b_L *if* a *and* b *are consecutive in* π*. We have also solid edges between* S *and* $(\pi_1)_L$ *and between* $(\pi_n)_R$ *and* T*. We have dashed edges between the vertices* $2i-1$ *and* $2i$*, for every* $i = 1, ..., n$ *and between* S *and* 1 *and between* $2n$ *and* T*.*

Note that each vertex in $G(\pi)$ is of degree 2, having exactly one solid edge and one dashed edge incident to it. Consequently, the breakpoint graph decomposes uniquely into disjoint alternating cycles, i.e. cycles in which the solid edges and the dashed edges alternate.

For a given cycle, we call its *length* the number of solid edges it contains, or equivalently, the number of dashed edges.

In the example from Fig. 3.1, for $\pi = [+3, -4, -2, +1, +5]$ the breakpoint graph $G(\pi)$ decomposes into two alternating cycles, one of length 1 and one of length 5.

For every signed permutation $\pi \in B_n$, the following lower bound holds :

$$d_{rev}(\pi, Id) \geq n + 1 - c(\pi), \tag{3.1}$$

where $c(\pi)$ denotes the number of alternating cycles in $G(\pi)$ (see Bafna and Pevzner [3]).

The bound (3.1) approximates the reversal distance extremely well for both simulated (see Kececioglu and Sankoff [26]) and biological data (see Bafna and Pevzner [3]). Kece-

cioglu and Sankoff [26] observed that the average difference between this bound and the exact distance is less than 1 for a random permutation.

Hannenhalli and Pevzner [20] proved that for every signed permutation $\pi \in B_n$ we have the exact formula

$$d_{rev}(\pi, Id) = n + 1 - c(\pi) + h(\pi) + f(\pi), \tag{3.2}$$

where $h(\pi)$ is the number of *hurdles* in $G(\pi)$ and $f(\pi)$ is 1 if π is a *fortress* and 0 otherwise.

For defining the notions of hurdle and fortress we need some more preliminaries.

A cycle is called *oriented* if it has length 1 or if, when we traverse it, we do not traverse all the solid edges in the same direction. Otherwise, the cycle is called *unoriented*.

We can define an equivalence relation on the cycles. An interval on a genome is a segment of consecutive genes. We say that two cycles are equivalent if every interval containing all the vertices of the first cycle intersect every interval containing all the vertices of the second cycle. The equivalence classes are called *components*. A component is called *oriented* if it contains at least one oriented cycle and *unoriented* otherwise.

If there is an interval that contains an unoriented component τ, but no other unoriented components, then τ is called a *hurdle*. If there is an interval which contains exactly two unoriented components, among which exactly one is a hurdle, then this hurdle is called a *super hurdle*. If the breakpoint graph $G(\pi)$ contains an odd number of hurdles and all of them are super hurdles, then π is called a *fortress*.

The exact formula (3.2) for the reversal distance leads to a polynomial-time algorithm for sorting signed permutations by reversals. The most efficient algorithm at present runs in $\mathcal{O}(n^2)$ time and is due to Kaplan, Shamir and Tarjan [25].

The problem of computing the reversal distance for signed permutations can be solved in $\mathcal{O}(n)$ time (see Bader, Moret and Yan [2]).

Caprara [8] showed that genomes containing hurdles are very rare. For example, less than one percent of the genomes with 8 genes contain hurdles and only one in 10^5 genomes with 100 genes.

We will thus use the bound (3.1) as an approximation for the reversal distance.

The goal of the present work is to find the distribution of $c(\pi)$ for a random signed permutation, which will further imply a good approximation for the distribution of the reversal distance.

This would allow us to use the reversal distance as a measure for the exceptionality of the order of the orthologs in conserved genomic regions, using the first idea of Section 2.3, when we compare only the order of the orthologs which are in common between the observed conserved genomic region and the reference region. The reversal distance is more biologically relevant as the mathematical transposition distance used in the Chapter 2. In return, calculating its distribution for a random permutation is much more difficult. For example, this prevents us from using the ideas in Sections 2.4 and 2.5, which were specifically adapted to the reference region approach.

3.1.2 Previous results

Recently, Doignon and Labarre [13] have found the exact distribution of the number of alternating cycles in the breakpoint graph of a random *unsigned* permutation.

Sankoff and Haque [34] use a constructive approach to obtain asymptotic estimates for the distribution of the number of cycles in the breakpoint graph of two random *signed* permutations. Xu, Zheng and Sankoff [35] use the same approach to treat the case of multichromosomal genomes.

But for signed permutations, the exact distribution of the number of cycles in the breakpoint graph is still unknown.

In this article we obtain this distribution in terms of a product of transition probability matrices of a certain finite Markov chain.

3.2 The distribution of $c(\pi)$

3.2.1 The Markov chain imbedding technique

We will use the finite Markov chain imbedding technique introduced by Fu and Koutras [16].

Let X be a bounded random variable taking its values in the set of non negative integers.

Definition 3.2. *The $\{0, 1, ..., \ell\}$- valued r.v. X is called* Markov chain imbeddable *if*

(*i*) *there exists a positive integer n,*

(*ii*) *there exists a (possibly non-homogeneous) finite Markov chain $\{Y_t : 1 \leq t \leq n\}$ with values in a finite state space $E = \{a_1, ..., a_m\}$,*

(*iii*) *there exists a finite partition $\{C_x, x = 0, 1, ..., \ell\}$ on E, and*

(*iv*) *for every $x = 0, 1, ..., \ell$ we have*

$$\mathbb{P}(X = x) = \mathbb{P}(Y_n \in C_x).$$

The distribution of X can in this case be obtained as a product of transition matrices of the Markov chain $(Y_t)_{1 \leq t \leq n}$. Indeed, if we define $\{P_t(x, y); x, y \in E\}$ by

$$P_t(x, y) = \mathbb{P}(Y_t = y | Y_{t-1} = x),$$

then we have

$$\mathbb{P}(X = x) = \mu_1 P_2 \times \cdots \times P_n \left(\sum_{i: a_i \in C_x} e_i \right), \tag{3.3}$$

where

$$\mu_1 = (\mathbb{P}(Y_1 = a_1), ..., \mathbb{P}(Y_1 = a_m))$$

is the row vector of the initial probability of the Markov chain and, for each $i = 1, ..., m$, e_i is the column vector having 1 at the i-th coordinate and 0 elsewhere.

3.2.2 Our results

Let n be a fixed positive integer.

In our case, the variable of interest is $X := c(\pi_n)$, for which we will show that it is Markov chain imbeddable.

We will construct a finite state Markov chain $(Y_t)_{1\leq t\leq n}$ as in Definition 3.2, as follows.

We start with π_1 being a random signed permutation with 1 element, hence $\pi_1 = [+1]$ with probability $1/2$ and $\pi_1 = [-1]$ with probability $1/2$.

For every $t = 2, ..., n$, we let π_t represent the random signed permutation of t elements which is obtained from π_{t-1} by inserting at random the element t uniformly into one of the t possible positions, with the "+" sign with probability $1/2$ and the "-" sign with probability $1/2$, the sign being independent of the position.

Note that $(\pi_t)_{1\leq t\leq n}$ is a non-homogeneous Markov chain with initial distribution

$$\mathbb{P}(\pi_1 = [1]) = \mathbb{P}(\pi_1 = [-1]) = 1/2$$

and the following transition probability matrices : for every $2 \leq t \leq n$,

$$M_t(\sigma, \sigma^{+,i}) = \mathbb{P}(\pi_t = \sigma^{+,i} | \pi_{t-1} = \sigma) = \frac{1}{2t},$$
$$M_t(\sigma, \sigma^{-,i}) = \mathbb{P}(\pi_t = \sigma^{-,i} | \pi_{t-1} = \sigma) = \frac{1}{2t},$$

where $\sigma^{+,i} := [\sigma_1, ..., \sigma_{i-1}, t, \sigma_i, ..., \sigma_{t-1}]$ and $\sigma^{-,i} := [\sigma_1, ..., \sigma_{i-1}, -t, \sigma_i, ..., \sigma_{t-1}]$ and $M_t(\sigma, \sigma') = 0$ for every other $\sigma' \in B_t$.

It is easy to see that for every $t = 1, ..., n$, π_t is a random signed permutation of t elements, uniformly chosen among the $2^t t!$ elements of B_t.

We are interested in the distribution of $c(\pi_n)$.

For every $t = 1, ..., n$, we denote by $K_{j,t}, j = 1, ..., n+1$ the r.v.'s representing the number of cycles of length j in the breakpoint graph of the permutation π_t. We also denote by L_t the length of the cycle in $G(\pi_t)$ which contains the terminal point T.

For every $t = 1, ..., n$ we obviously have $K_{j,t} = 0$ for $j = t+2, ..., n+1$ and

$$\sum_{j=1}^{t+1} jK_{j,t} = t+1,$$
$$\sum_{j=1}^{t+1} K_{j,t} = c(\pi_t).$$

We let

$$Y_t := (L_t, K_{1,t}, ..., K_{n+1,t}), t = 1, ..., n.$$

We will call Y_t *the type* of the permutation π_t.

For example, the permutation $\pi = [+3, -4, -2, +1, +5]$ from Fig. 3.1 is of type $(\ell, \overrightarrow{k})$, where $\ell = 1$, $\overrightarrow{k} = (1, 0, 0, 0, 1, 0)$.

Note that for every $1 \leq t \leq n$, Y_t takes values in the finite set

$$E_t = \left\{ (\ell, k_1, ..., k_{t+1}, \underbrace{0, ..., 0}_{n-t}) : \ell \in \{1, ..., t+1\}, \sum_{j=1}^{t+1} jk_j = t+1, k_\ell \geq 1 \right\}.$$

We have

$$\mathbb{P}(c(\pi_n) = x) = \mathbb{P}(Y_n \in C_x),$$

where, for every $x = 1, 2, ..., n+1$:

$$C_x = \left\{ (\ell, k_1, ..., k_{n+1}) : \sum_{j=1}^{n+1} k_j = x, \sum_{j=1}^{n+1} j k_j = n+1, k_\ell \geq 1 \right\}.$$

We will show that $(Y_t)_{1 \leq t \leq n}$ is a non-homogeneous Markov chain.

The initial distribution of Y_1 is

$$\mathbb{P}(Y_1 = (1, 2, 0, 0, ..., 0)) = \mathbb{P}(Y_1 = (2, 0, 1, 0, ..., 0)) = 1/2,$$

the first case $Y_1 = (1, 2, 0, 0, ..., 0)$ corresponding to $\pi_1 = [+1]$ and the second case $Y_1 = (2, 0, 1, 0, ..., 0)$ corresponding to $\pi_1 = [-1]$.

Let us write $\overrightarrow{k} := (k_1, ..., k_{n+1})$.

For $2 \leq t < n$, write $Y_{t-1} = (\ell, \overrightarrow{k})$. Note that necessarily $k_\ell \geq 1$.

We have the following result.

Proposition 3.1. *$(Y_t)_{1 \leq t \leq n}$ is a non-homogeneous Markov chain of initial distribution*

$$\mathbb{P}(Y_1 = (1, 2, 0, 0, ..., 0)) = \mathbb{P}(Y_1 = (2, 0, 1, 0, ..., 0)) = 1/2,$$

and of transition probabilities described as follows. If $Y_{t-1} = (\ell, \overrightarrow{k})$, with $k_\ell \geq 1$, then the possible transitions are to $Y_t = (\ell', \overrightarrow{k'})$, where

(i) $\ell' = \ell + 1$ and $\overrightarrow{k'} = \overrightarrow{k} - e'_\ell + e'_{\ell+1}$, with probability $\ell/(2t)$;

(ii) $\ell' = j$, with $1 \leq j \leq \ell$, and $\overrightarrow{k'} = \overrightarrow{k} - e'_\ell + e'_j + e'_{\ell+1-j}$, with probability $1/(2t)$;

(iii) $\ell' = \ell + x + 1$, with $1 \leq x \leq t - \ell, x \neq \ell$ and $\overrightarrow{k'} = \overrightarrow{k} - e'_\ell - e'_x + e'_{\ell+x+1}$, with probability xk_x/t;

(iv) $\ell' = 2\ell + 1$ and $\overrightarrow{k'} = \overrightarrow{k} - 2e'_\ell + e'_{2\ell+1}$, with probability $\ell(k_\ell - 1)/t$,

where for each i, e'_i is the row vector having 1 at the i-th coordinate and 0 elsewhere.

Proof. We will show that $\mathbb{P}(Y_t = (\ell', \overrightarrow{k'}) | \pi_{t-1} = \pi)$, where π is a permutation of type $(\ell, \overrightarrow{k})$, depends only on $\ell', \overrightarrow{k'}, \ell, \overrightarrow{k}$. This will imply that $\mathbb{E}[f(Y_t)|\pi_{t-1}]$ is Y_{t-1}- measurable, for every bounded measurable function $f : E_t \to \mathbb{R}$. Then, by the fact that $(\pi_t)_{1 \leq t \leq n}$ is a Markov chain, it will follow that $(Y_t)_{1 \leq t \leq n}$ is also a Markov chain.

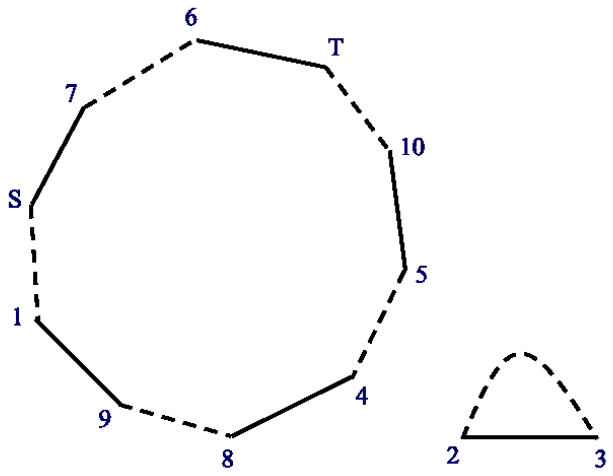
Suppose now that $\pi_{t-1} = \pi$, with π being of type $(\ell, \overrightarrow{k})$.

In Fig. 3.2 we have the disjoint cycle decomposition of the breakpoint graph of the permutation $\pi_5 = \pi = [+4, -2, -1, +5, +3]$. In this case we have $\ell = 5$, $\overrightarrow{k} = (1, 0, 0, 0, 1, 0)$.

We will investigate the changes produced in the breakpoint graph when inserting the new element $\pm t$, at random, into one of the t possible positions of the permutation π, with the "+" sign with probability $1/2$ and the "$-$" sign with probability $1/2$.

The modifications concerning the dashed edges are simple. Disregarding the sign of $\pm t$, the dashed edge between $2(t-1)$ and T is deleted and replaced by a dashed edge between

FIG. 3.2 – The disjoint cycle decomposition of $G(\pi)$, for $\pi = [+4, -2, -1, +5, +3]$.



$2(t-1)$ and $2t-1$, and then another dashed edge is added between $2t$ and T (see for example Fig. 3.3).

Concerning the solid edges : we choose at random a solid edge among the t solid edges in the breakpoint graph of π, we delete it and then add two other solid edges to connect the two extremities of the deleted edge to $2t-1$ and $2t$ respectively, in one of the two possible ways. The choice of the solid edge to be deleted corresponds to the choice of the position in the permutation π where $\pm t$ is inserted. The way in which we connect the two extremities of the deleted edge to $2t-1$ and $2t$ respectively, corresponds to the sign of the element t.

More precisely, if we choose to insert the element $\pm t$ in the position i, where $2 \leq i \leq t-1$, then we will delete the solid edge between $(\pi(i-1))_R$ and $(\pi(i))_L$ (see the notations in Definition 3.1). If we insert $+t$, then we will add two solid edges between $(\pi(i-1))_R$ and $2t-1$ and between $2t$ and $(\pi(i))_L$. If we insert $-t$, then we will add two solid edges between $(\pi(i-1))_R$ and $2t$ and between $2t-1$ and $(\pi(i))_L$.

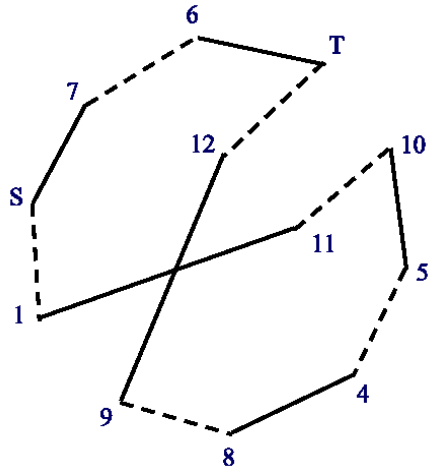
If we choose to insert the element $\pm t$ in the position 1, i.e. at the beginning of the permutation π, then we will delete the solid edge between S and $(\pi(1))_L$. If we insert $+t$ we add two solid edges between S and $2t-1$ and between $2t$ and $(\pi(1))_L$, and if we insert $-t$ we add two solid edges between S and $2t$, and between $2t-1$ and $(\pi(1))_L$.

If we choose to insert $\pm t$ in the position t, i.e. at the end of the permutation π, then we will delete the solid edge between $(\pi(t))_R$ and T. If we insert $+t$ we add two solid edges between $(\pi(t-1))_R$ and $2t-1$ and between $2t$ and T, and if we insert $-t$ we add two solid edges between $(\pi(t-1))_R$ and $2t$ and between $2t-1$ and T.

The cycle structure of the breakpoint graph will change as follows.

If we delete a solid edge belonging to the cycle of size ℓ which contains T, then we have two possible situations, depending on the deleted solid edge and on the permutation π.

FIG. 3.3 – For $\pi_6 = [+4, -2, -1, +6, +5, +3]$ the cycle containing T grows to the length $\ell' = \ell + 1 = 6$. The element $+6$ is inserted into $\pi_5 = [+4, -2, -1, +5, +3]$ in position $i = 4$, corresponding to the deletion of the solid edge 1–9.



One possible situation is that, when we insert $+t$, the cycle containing T grows to the length $\ell + 1$ (see Fig. 3.3), and when we insert $-t$ it splits into two smaller cycles, of sizes which sum to $\ell + 1$ (see Fig. 3.4).

The other possible situation is the converse, i.e. when we insert $-t$ the cycle containing T becomes of size $\ell + 1$ (see Fig. 3.5), and when we insert $+t$ it splits into two smaller cycles, of sizes which sum to $\ell + 1$ (see Fig. 3.6).

The event that the cycle containing T becomes of size $\ell + 1$ occurs with probability $\ell/(2t)$, because we have ℓ possible solid edges to choose in the cycle containing T. In the case when the cycle containing T splits into two cycles, the new size j of the cycle which will contain T is chosen at random, uniformly between 1 and ℓ. The size of the second cycle is then simply $\ell + 1 - j$. Each size j corresponds to a specific choice for the deleted solid edge, hence the event that the cycle containing T splits into two cycles and the size of the new cycle which will contain T becomes j, occurs with probability $1/(2t)$.

The cases (i) and (ii) in the statement correspond to the deletion of a solid edge from the cycle containing T, and the cases (iii) and (iv) correspond to the deletion of a solid edge belonging to a cycle not containing T.

If we delete a solid edge from a cycle not containing T, then, disregarding the sign of t, this cycle will merge with the one containing T. If the cycle from which we have deleted a solid edge was of size x, then in the breakpoint graph of π_t the cycle containing T will be of size $\ell + x + 1$.

In (iii), x represents the length of the cycle not containing T from which we choose a solid edge to be deleted. If $x \neq \ell$, the probability that this event occurs equals xk_x/t, because we have k_x cycles of length x that we can choose, and each of them contains x

FIG. 3.4 – For $\pi_6 = [+4, -2, -1, -6, +5, +3]$ the cycle containing T splits into two and $\ell' = 3$. The element -6 is inserted into $\pi_5 = [+4, -2, -1, +5, +3]$ in position $i = 4$, corresponding to the deletion of the solid edge 1–9.

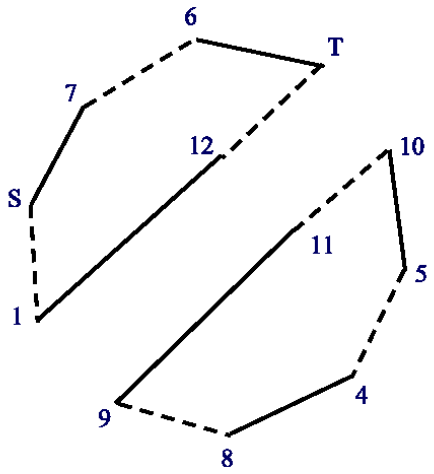


FIG. 3.5 – For $\pi_6 = [+6, +4, -2, -1, +5, +3]$ the cycle containing T splits into two and $\ell' = 2$. The element $+6$ is inserted into $\pi_5 = [+4, -2, -1, +5, +3]$ in position $i = 1$, corresponding to the deletion of the solid edge S–7.

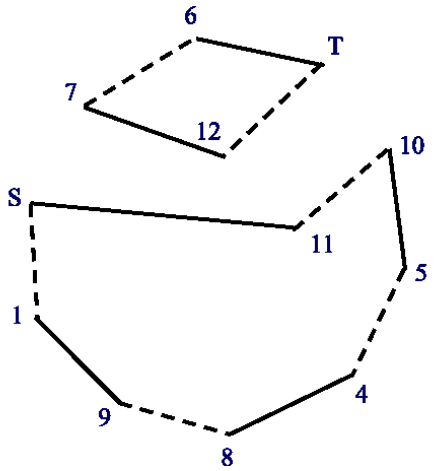
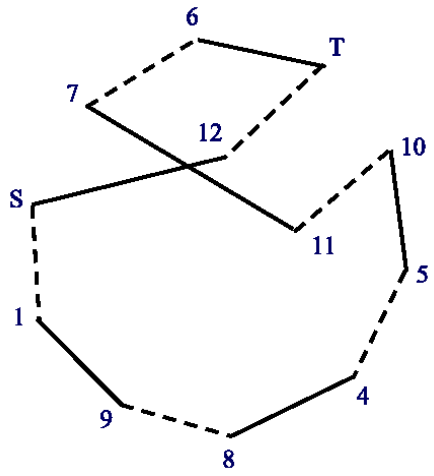


FIG. 3.6 – For $\pi_6 = [-6, +4, -2, -1, +5, +3]$ the cycle containing T grows to the length $\ell' = \ell + 1 = 6$. The element -6 is inserted into $\pi_5 = [+4, -2, -1, +5, +3]$ in position $i = 1$, corresponding to the deletion of the solid edge S–7.



solid edges.

The case (iv) corresponds to the case $x = \ell$, when we have only $k_\ell - 1$ possibilities to choose a cycle of size ℓ not containing T. □

It can be easily checked that the sum of all the transition probabilities given in Proposition 3.1 equals 1.

Proposition 3.1 describes the entries of the transition probability matrix P_t of the non-homogeneous Markov chain $(Y_t)_t$. As described in subsection 3.1, formula (3.3), we can therefore obtain the distribution of $c(\pi_n)$ as the product of n transition matrices.

3.3 Concluding remarks

In this chapter we have obtained the distribution of the number of alternating cycles in the breakpoint graph of a random signed permutation, in the form of a product of transition probability matrices of a certain finite Markov chain, using the Markov chain imbedding technique.

A drawback of our method is the fact that our Markov chain is non-homogeneous and of large dimension, inducing a high computational complexity.

We have implemented an iterative procedure which, for a given n, computes numerically the distribution of Y_n and then that of $c(\pi_n)$. At each step $t = 1, ..., n-1$ we compute the distribution of Y_{t+1} from the distribution of Y_t, using the transition probabilities described in Proposition 3.1.

TAB. 3.1 – The distribution of $c(\pi)$ for a random $\pi \in B_{20}$.

k	1	2	3	4	5	6	7	8	9
p_k	0.19213	0.34805	0.27688	0.13047	0.04126	0.00938	0.00160	0.00021	0.00002

TAB. 3.2 – The distribution of $c(\pi)$ for a random $\pi \in B_{30}$.

k	1	2	3	4	5	6	7	8	9	10
p_k	0.15849	0.31791	0.28690	0.15704	0.05909	0.01639	0.0035	0.00059	0.00008	0.00001

The complexity of our algorithm is of the order $n^2 \times p(n+1)$, where p is the partition function, i.e. for every positive integer m, $p(m)$ is the number of integer partitions of m. An asymptotic expression for $p(m)$ is given by

$$p(m) \sim \frac{\exp(\pi\sqrt{(2m)/3})}{4m\sqrt{3}}, \text{ as } m \to \infty.$$

In the tables Tab. 3.1 and Tab. 3.2 we give the distribution of $c(\pi)$ for a random uniform signed permutation of 20 and 30 elements, respectively. In the two tables, p_k denotes the probability that $c(\pi)$ takes the value k. For the k's which do not appear in the table, the corresponding probability is negligible.

On a Pentium 4 processor, 3.1 Mhz, 512 Mb, the computation time was $13s$ for $n = 20$, $300s$ for $n = 30$, $4 \times 10^3 s$ for $n = 40$ and $4 \times 10^4 s$ for $n = 50$.

We have not succeeded yet to solve analytically the recursive relations linking the distribution of Y_t to the distribution of Y_{t-1}.

A plan for a future work is to try to find a closed analytic formula for the exact distribution of the number of cycles in the breakpoint graph of a random signed permutation.

Chapitre 4

Applications to biological data

This chapter is devoted to some applications of our results to real biological data.

4.1 Comparison Homo-Sapiens and Oryzias-Latipes

In this example we are interested in finding signs for the conservation of the Major Histocompatibility Complex (MHC) between the human genome and the genome of Oryzias-Latipes (or Japanese killifish, a very small ricefish, popular as an aquarium fish native to Southeast Asia).

The Major Histocompatibility Complex contains genes involved in the immune defense. In the human genome, as the result of two rounds of polyploidization (whole genome duplication), we find four MHC paralogous regions (see [1] for more informations on the Major Histocompatibility Complex).

We choose as reference region for our analysis the MHC paralogous region on the human chromosome 9 (129045207–140191570). The numbers in brackets represent the positions, on the chromosome 9, of the starting and, respectively, the ending nucleotide of the region. It has been shown that this region evolves slower than the other three.

This region contains 38 genes which have at least one ortholog in the genome of Oryzias-Latipes. Among those 38 genes, 8 have two orthologs in Oryzias-Latipes and 30 have one single ortholog.

Therefore, using the notations from Chapter 1, the data are the following :

- $m = 38$: the number of genes in the reference region in the human genome (the species A) which have at least one ortholog in the genome of Oryzias-Latipes (the species B) ;
- $\phi' = [1, 2]$: the vector containing all the distinct values for the sizes of the multigene families in the Oryzias-Latipes genome ;
- $g = [30, 8]$: the vector containing the multiplicities of the different sizes in ϕ' ;
- $n = 46$: the total number of genes in Oryzias-Latipes which are orthologous of genes in the human reference region ;
- $N = 19686$: the size of the genome of Oryzias-Latipes (the total number of genes).

After localizing the 46 orthologs in the genome of Oryzias-Latipes, 9 potential conserved genomic regions were identified : 3 regions on the chromosome 9 and the six others on the chromosome 12 (see Fig. 4.1 and Fig. 4.2).

For each of those regions we determine its weight h and its length r (in the simplified model of the genome B seen as $[0,1]$), and then we calculate its p-value using the compound Poisson approximation from Chapter 1, as described in subsubsection 1.4.3.

The results are as follows.

Region #1 : chromosome 9 (899561–1206257), cluster Id : 35 in Fig. 4.1
– contains 3 orthologs, of labels $\frac{1}{2}, \frac{1}{2}, \frac{1}{2}$;
– $h = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} = 1.5$;
– the total number of genes in the region : 9 ;
– $r = \frac{9}{19686}$;
– p-value : 0.5636 (not significant).

Region #2 : chromosome 9 (28437906–29203467), cluster Id : 152 in Fig. 4.1
– contains 4 orthologs, of labels $1, \frac{1}{2}, \frac{1}{2}, 1$;
– $h = 1 + \frac{1}{2} + \frac{1}{2} + 1 = 3$;
– the total number of genes in the region : 22 ;
– $r = \frac{22}{19686}$;
– p-value : 0.0148 (significant at the level $\alpha = 0.05$).

Region #3 : chromosome 9 (31902437–32170260), cluster Id : 185 in Fig. 4.1
– contains 3 orthologs, of labels $\frac{1}{2}, \frac{1}{2}, 1$;
– $h = \frac{1}{2} + \frac{1}{2} + 1 = 2$;
– total number of genes in the region : 3 ;
– $r = \frac{3}{19686}$;
– p-value : 0.2985 (not significant).

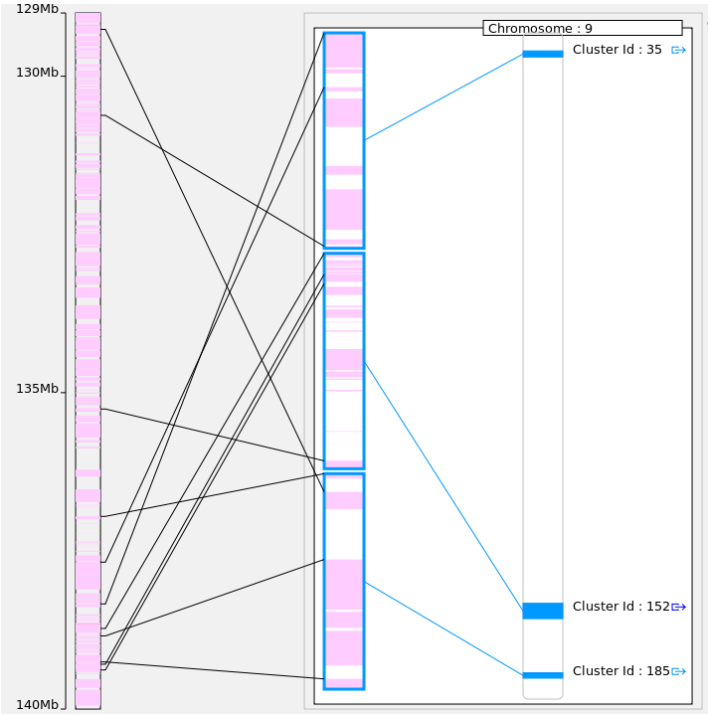
Region #4 : chromosome 12 (993203–5399518), cluster Id : 211 in Fig. 4.2
– contains 4 orthologs, of labels $1, 1, 1, \frac{1}{2}$;
– $h = 1 + 1 + 1 + \frac{1}{2} = 3.5$;
– total number of genes in the region : 7 ;
– $r = \frac{7}{19686}$;
– p-value : 1.63×10^{-5} (highly significant).

Region #5 : chromosome 12 (6945906–8246163), cluster Id : 235 in Fig. 4.2
– contains 6 orthologs, of labels $1, \frac{1}{2}, 1, 1, \frac{1}{2}, \frac{1}{2}$;
– $h = 1 + \frac{1}{2} + 1 + 1 + \frac{1}{2} + \frac{1}{2} = 4.5$;
– total number of genes in the region : 52 ;
– $r = \frac{52}{19686}$;
– p-value : 1.27×10^{-4} (very significant)

Region #6 : chromosome 12 (10049683–10113348), cluster Id : 288 in Fig. 4.2
– contains 3 orthologs, of labels $1, 1, 1$;
– $h = 1 + 1 + 1 = 3$;
– total number of genes in the region : 3 ;
– $r = \frac{3}{19686}$;
– p-value : 2.82×10^{-4} (very significant).

Region #7 : chromosome 12 (11625364–11880175), cluster Id : 332 in Fig. 4.2

FIG. 4.1 – The three regions on the chromosome 9 of Oryzias-Latipes.



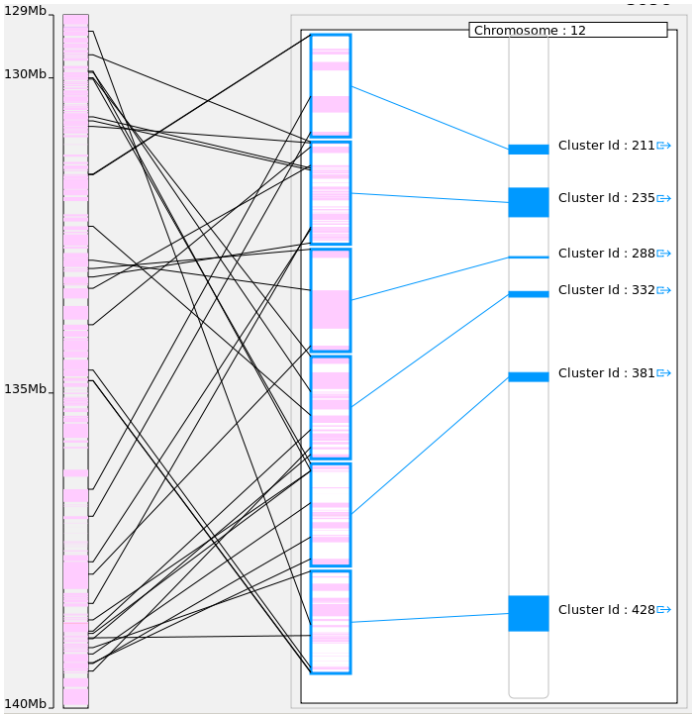
– contains 4 orthologs, of labels $1, 1, 1, 1$;
– $h = 1 + 1 + 1 + 1 = 4$;
– total number of genes in the region : 16 ;
– $r = \frac{16}{19686}$;
– p-value : 5.83×10^{-5} (highly significant).

Region #8 : chromosome 12 (15301431–15697269), cluster Id : 381 in Fig. 4.2
– contains 4 orthologs, of labels $1, 1, 1, \frac{1}{2}$;
– $h = 1 + 1 + 1 + \frac{1}{2} = 3.5$;
– total number of genes in the region : 18 ;
– $r = \frac{18}{19686}$;
– p-value : 2.71×10^{-4} (very significant).

Region #9 : chromosome 12 (25421295–26996650), cluster Id : 428 in Fig. 4.2
– contains 6 orthologs, of labels $\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 1, 1$;
– $h = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + 1 + 1 = 4$;
– total number of genes in the region : 30 ;
– $r = \frac{30}{19686}$;
– p-value : 3.81×10^{-4} (very significant).

The results indicate a high conservation between the human MHC region on the chromosome 9 and the six regions on the chromosome 12 of Oryzias-Latipes. The conservation

FIG. 4.2 – The six regions on the chromosome 12 of Oryzias-Latipes.



between the same human MHC region and the chromosome 9 of Oryzias-Latipes is less evident.

4.2 Comparison Ciona-Intestinalis and Homo-Sapiens

In this second analysis we will compare the human genome and the genome of Ciona-Intestinalis, which is a Urochordata (sea squirt) whose genome has been sequenced and which has become, over the past decade, a major experimental model for developmental biologists.

We start with a reference genomic region in Ciona, spread over chromosomes 4 and 10 and containing genes of the immunoglobulin superfamily.

The concatenated reference region contains 14 genes having at least one ortholog in the human genome.

With the notations from Chapter 1, the data are the following :

- $m = 14$: the number of genes in the reference region in Ciona which have at least one ortholog in the human genome ;
- $\phi' = [1, 2, 3, 4, 7, 8, 16]$: the vector containing all the distinct values for the sizes of the multigene families in the human genome ;
- $g = [5, 3, 2, 1, 1, 1, 1]$: the vector containing the multiplicities of the different sizes in ϕ' ;
- $n = 52$: the total number of genes in the human genome which are orthologous of genes in the reference region in Ciona ;
- $N = 36396$: the size of the human genome.

After localizing the 52 orthologs in the human genome, we found 5 potential conserved genomic regions, on the chromosomes 1, 3, 11, 19, 21 ; the regions on the chromosomes 19 and 21 will be analyzed both separated and considered as a single region (see Fig. 4.3).

The results are the following.

Region #1 : chromosome 1 (155749791–165754456)
- contains 9 orthologs, of labels $\frac{1}{3}, 2 \times \frac{1}{7}, 4 \times \frac{1}{8}, 2 \times \frac{1}{16}$;
- $h = 1.244$ (the weight of the region) ;
- the total number of genes in the region : 596 ;
- $r = \frac{596}{36396}$;
- p-value : 0.9798 (not significant).

Region #2 : chromosome 3 (106568403–123322672)
- contains 4 orthologs, of labels $\frac{1}{2}, \frac{1}{3}, \frac{1}{7}, \frac{1}{16}$;
- $h = 1.0387$;
- the total number of genes in the region : 140 ;
- $r = \frac{140}{36396}$;
- p-value : 0.8389 (not significant).

Region #3 : chromosome 11 (60495750–133526846)
- contains 14 orthologs, of labels $2 \times 1, 3 \times \frac{1}{2}, \frac{1}{3}, 2 \times \frac{1}{7}, 6 \times \frac{1}{16}$;
- $h = 3.494$;
- the total number of genes in the region : 998 ;
- $r = \frac{998}{36396}$;
- p-value : 0.0083 (significant at the level $\alpha = 0.01$).

Region #4 : chromosome 19 (40511919–60093650)

– contains 13 orthologs, of labels $3 \times 1, \frac{1}{2}, 2 \times \frac{1}{3}, \frac{1}{4}, 2 \times \frac{1}{7}, 4 \times \frac{1}{8}$;
– $h = 5.2024$;
– the total number of genes in the region : 803 ;
– $r = \frac{803}{36396}$;
– p-value : 1.7612×10^{-6} (highly significant).

Region #5 : chromosome 21 (17807201–44485277)
– contains 4 orthologs, of labels $\frac{1}{2}, 3 \times \frac{1}{16}$;
– $h = 0.6875$;
– the total number of genes in the region : 320 ;
– $r = \frac{320}{36396}$;
– p-value : 0.9999 (not significant).

Region #4+#5 : chromosomes 19 + 21
– contains 17 orthologs, of labels $3 \times 1, 2 \times \frac{1}{2}, 2 \times \frac{1}{3}, \frac{1}{4}, 2 \times \frac{1}{7}, 4 \times \frac{1}{8}, 3 \times \frac{1}{16}$;
– $h = 5.8899$;
– the total number of genes in the region : 1123 ;
– $r = \frac{1123}{36396}$;
– p-value : 5.0717×10^{-7} (highly significant).

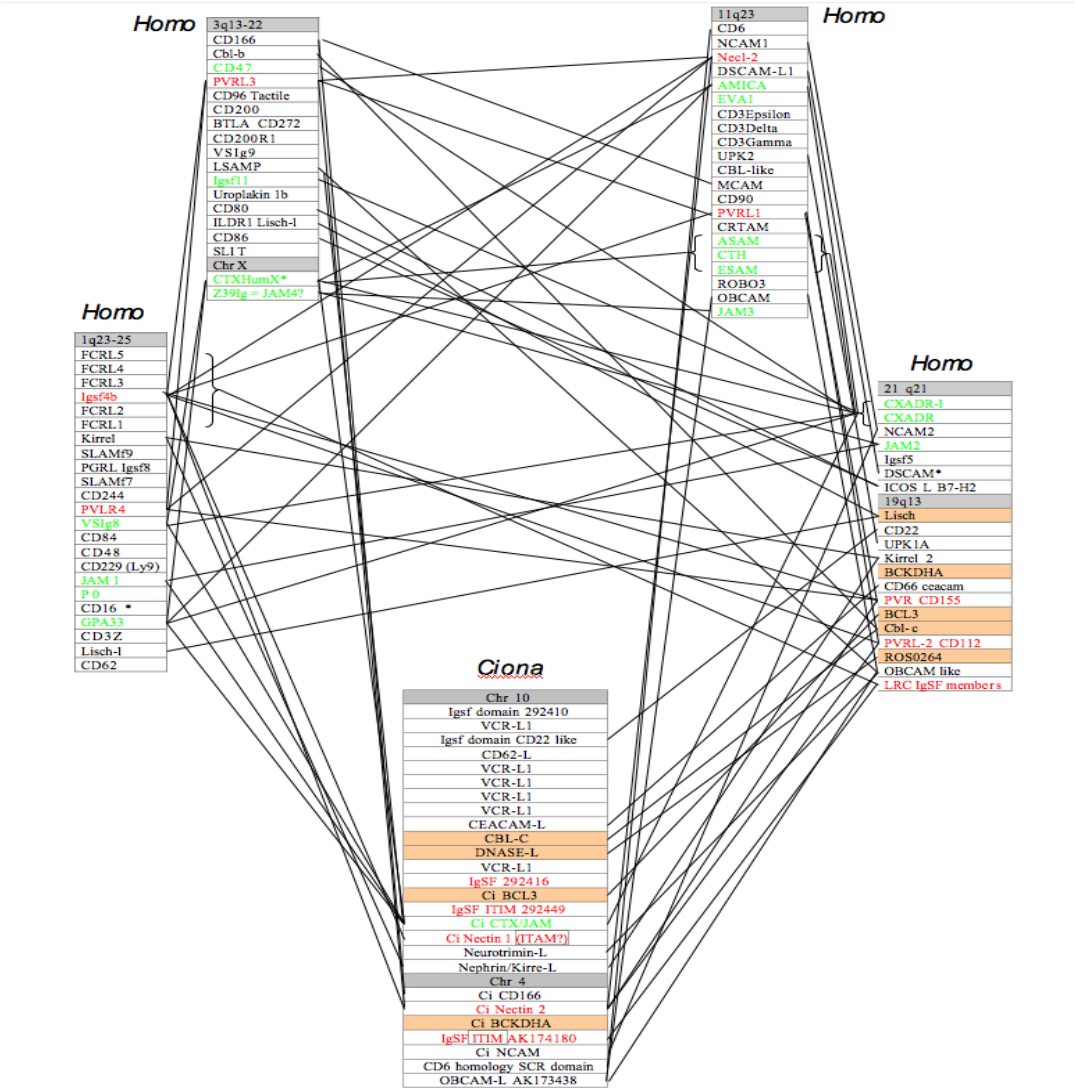
4.3 Concluding remarks

The data for the first example have been given to me by Pierre Pontarotti. The data for the second example have been given to me by Louis Pasquier (Basel University).

In the example Ciona-Intestinalis – Homo-Sapiens, the fact that our approach characterizes the regions numbers 1, 2 and 5 as non significant was a surprise to Louis Pasquier. Our test seems to be more severe than the other methods which have selected those regions.

Note that our way of weighting each gene from a multigenic family by the inverse of the family size is questionable. We shall implement other versions of our test on these examples, and will rediscuss with the biologists in a near future.

FIG. 4.3 – The orthology relationships between the four (five) regions in Homo-Sapiens and the concatenated reference region in Ciona-Intestinalis.



Conclusion et perspectives

Nous avons abordé dans cette thèse un sujet important dans le domaine de la génomique comparative, celui de la détection de régions génomiques conservées entre espèces.

Plus précisément, nous nous sommes intéressé à évaluer, du point de vue statistique, la significativité des régions génomiques conservées observées entre deux espèces différentes. Nous traitons le cas des régions génomiques conservées trouvées par une approche de type région de référence.

Dans un premier temps nous considérons seulement la proximité des orthologues dans ces régions génomiques conservées, et non pas leur ordre. Dans le Chapitre 1 nous utilisons une approximation de Poisson composée pour calculer la p-valeur d'une région génomique conservée donnée. Nous estimons l'erreur de notre approximation à l'aide de la méthode de Stein-Chen pour l'approximation de Poisson composée, l'approche par couplage.

Dans notre approche nous prenons en compte l'existence de familles multigéniques, en pondérant les orthologues en proportion inverse des tailles des familles multigéniques correspondantes.

Dans un travail futur nous envisageons d'essayer d'autres façons de prendre en compte les orthologues multiples. Par exemple, une autre idée possible sera de compter, pour un cluster donné, le nombre de familles multigéniques différentes qui sont représentées dans ce cluster. Cette idée est utilisée par Raghupathy et Durand (voir [29]) dans le cas des clusters trouvés par une approche de type "window-sampling".

Un autre problème que nous espérons résoudre dans un travail futur est celui des tests multiples causé par le fait que l'on ne fixe pas a priori la taille des clusters orthologues cherchés.

Dans la deuxième partie de la thèse, nous nous sommes intéressés à l'ordre des orthologues dans les régions génomiques conservées. Nous traitons seulement le cas sans famille multigénique, i.e. nous supposons que chaque gène de la région de référence a au plus un orthologue dans le deuxième génome.

Dans le Chapitre 2 nous avons proposé trois mesures, basées sur la distance de transposition dans le groupe symétrique, pour mesurer l'exceptionalité de l'ordre des gènes dans une région génomique conservée donnée. Nous avons obtenu explicitement la distribution de ces mesures pour une permutation aléatoire, i.e. sous l'hypothèse nulle de l'ordre aléatoire des gènes dans le génome.

Dans le troisième chapitre de la thèse nous traitons le cas des permutations signées, qui correspond aux situations biologiques dans lesquelles on connaît aussi l'orientation des gènes dans les génomes. Notre but a été de trouver la loi du nombre de cycles dans le graphe des points de rupture d'une permutation signée aléatoire. La connaissance de cette loi fournit, par la suite, une bonne approximation de la loi de la distance d'inversion pour

une permutation signée aléatoire. Ceci nous permet d'utiliser la distance d'inversion comme mesure de l'exceptionalité de l'ordre des gènes dans des régions génomiques conservées, en utilisant l'idée de la Section 2.3.

Nous avons obtenu la loi du nombre de cycles dans le graphe des points de rupture d'une permutation signée aléatoire en utilisant la méthode "Markov chain imbedding". Nous avons construit une chaîne de Markov finie non-homogène qui nous permet d'exprimer la loi du nombre de cycles dans le graphe des points de rupture comme un produit de matrices de transition de cette chaîne. Un point faible de notre méthode est la grande dimension de cette chaîne de Markov, qui entraîne une grande complexité des calculs.

Trouver une formule explicite pour la loi du nombre de cycles dans le graphe des points de rupture d'une permutation signée aléatoire reste un problème ouvert.

Un autre problème que nous n'avons pas réussi à résoudre dans cette thèse est la comparaison de l'ordre des gènes dans le cas avec familles multigéniques, qui revient à la comparaison de multipermutations.

Nous pouvons essayer aussi de généraliser les idées du Chapitre 2 à d'autres distances, peut-être plus pertinentes du point du vue biologique que la distance de transposition.

Une autre question à laquelle nous n'avons pas encore trouvé une réponse satisfaisante est comment combiner, dans un seul test statistique, la prise en compte de la proximité des orthologues et la prise en compte de leur ordre.

Une ouverture encore plus générale est la comparaison de plusieurs espèces différentes à la fois, le grand défi étant la reconstruction des génomes ancestraux.

Bibliographie

[1] Abi-Rached, L., Gilles, A., Shiina, T., Pontarotti, P. and Inoko, H. (2002). Evidence of en bloc duplication in vertebrate genomes. *Nat. Genetics* **31**, 100–105.

[2] Bader, D. A., Moret B. M. E. and Yan, M. (2001). A linear-time algorithm for computing inversion distance between signed permutations with an experimental study. *J. Comput. Biology* **8**, 483–491.

[3] Bafna, V. and Pevzner, P. (1996). Genome rearrangements and sorting by reversals. *SIAM J. Comput.* **25**, 272–289.

[4] Barbour, A., Chen, L. and Loh, W. (1992). Compound Poisson approximation for nonnegative random variables via Stein's method. *The Annals of Probability* **20**, 1843–1866.

[5] Barbour, A. and Xia, A. (2000). Estimating Stein's constants for compound Poisson approximation. *Bernoulli* **6**, 581–590.

[6] Bergeron, A., Corteel, S. and Raffinot, M. (2002). The algorithmic of gene teams. *Lecture Notes in Computer Science* **2452**, 464–476.

[7] Bollobas, B. (2001). *Random Graphs*. Cambridge University Press.

[8] Caprara, A. (1999). On the tightness of the alternating-cycle lower bound for sorting by reversals. *J. Combin. Optim.* **3**, 149–182.

[9] Christie, D. A. (1996). Sorting permutations by block-interchanges. *Information Processing Letters* **60**, 165–169.

[10] Comtet, L. (1974). *Advanced Combinatorics : The Art of Finite and Infinite Expansions*. Reidel Publishing Company, Dordrecht-Holland/Boston-U.S.A.

[11] Danchin, E. and Pontarotti, P. (2004). Statistical evidence for a more than 800-million-year-old evolutionarily conserved genomic region in our genome. *J. Mol. Evol.* **59**, 587–597.

[12] Diaconis, P. (1988). *Group Representations in Probability and Statistics*. Inst. Math. Stat., Hayward.

[13] Doignon, J-P. and Labarre, A. (2007). On Hultman numbers. *Journal of Integer Sequences* **10**, Article 07.6.2.

[14] Durand, D. and Sankoff, D. (2003). Tests for gene clustering. *J. Comput. Biology* **10**, 453–482.

[15] Eriksen, N. and Hultman, A. (2004). Estimating the expected reversal distance after a fixed number of reversals. *Advances in Applied Mathematics* **32**, 439–453.

[16] Fu, J. C. and Koutras, M. V. (1994). Distribution theory of runs : a Markov chain approach. *J. Amer. Statist. Soc.* **89**, 1050–1058.

[17] Glaz, J and Naus, J. (1983). Multiple clusters on the line. *Commun. Statist. - Theor. Meth.* **12**, 1961–1986.

[18] Glaz, J, Naus, J., Roos, M. and Wallenstein, S. (1994). Poisson approximations for the distribution and moments of ordered m- spacings. *J. Appl. Probab.* **31**, 271–281.

[19] Glaz, J., Naus, J. and Wallenstein, S. (2001). *Scan Statistics*. Springer Verlag, New York.

[20] Hannenhalli, S. and Pevzner, P. (1995). Transforming cabbage into turnip : polynomial algorithm for sorting signed permutations by reversals. *Proc. 27th Annual ACM Symposium on the Theory of Computing*, ACM Press, New York, 178–189.

[21] Heber, S. and Stoye, J. (2001). Finding all commun intervals of k permutations. *In Proceedings of CPM01, Lecture Notes in Computer Science*, **2089**, 207–218, Springer Verlag, Berlin.

[22] Hoberman, R. and Durand, D. (2005). The incompatible desiderata of gene cluster properties. *Lecture Notes in Bioinformatics* **3678**, 73–87.

[23] Hoberman, R., Sankoff, D. and Durand, D. (2005). The statistical analysis of spatially clustered genes under the maximum gap criterion. *Journal of Computational Biology* **12**, 1083–1102.

[24] Huntington, R. J. and Naus, J. (1975). A simpler expression for kth nearest neighbor coincidence probabilities. *The Annals of Probability* **3**, 894–896.

[25] Kaplan, H., Shamir, R. and Tarjan, R. E. (1997). Faster and simpler algorithm for sorting signed permutations by reversals. *Proc. 8th Annual ACM-SIAM Symposium on Discrete Algorithms*, ACM Press, New York, 344–351.

[26] Kececioglu, J. and Sankoff, D. (1995). Exact and approximation algorithms for sorting by reversals, with application to genome rearrangement. *Algorithmica* **13**, 180–210.

[27] Li, Z., Wang, L. and Zhang, K. (2006). Algorithmic approaches for genome rearrangement : a review. *IEEE Transactions on Systems, Man and Cybernetics* Part C, **36**, 636–648.

[28] Naus, J. (1974). Probabilities for a generalized birthday problem. *Journal of the American Statistical Association* **69**, 810–815.

[29] Raghupathy, N. and Durand, D. (2005). Individual gene cluster statistics in noisy maps. *Lecture Notes in Bioinformatics* **3678**, 106–120.

[30] Roos, M. (1993). Stein-Chen method for compound Poisson approximation. PhD Thesis, Univ. Zurich.

[31] Roos, M. (1993). Compound Poisson approximations for the number of extreme spacings. *Adv. Appl. Probab.* **25**, 847–874.

[32] Pyke, R. (1965). Spacings. *J. R. Statist. Soc. B* **27**, 395–449.

[33] Sankoff, D. and Haque, L. (2005). Power boosts for clusters tests. *Lecture Notes in Computer Science* **3678**, 121–130.

[34] Sankoff, D. and Haque, L. (2006). The distribution of genomic distance between random genomes. *Journal of Computational Biology* **13**, 1005–1012.

[35] Xu, W., Zheng, C. and Sankoff, D. (2006). Paths and cycles in breakpoint graphs of random multichromosomal genomes. *Lecture Notes in Bioinformatics* **4205**, 51–62.

Résumé

Cette thèse se concentre sur quelques sujets de probabilités et statistique liés à la génomique comparative. Dans la première partie nous présentons une approximation de Poisson composée pour calculer des probabilités impliquées dans des tests statistiques pour la significativité des régions génomiques conservées trouvées par une approche de type région de référence. Un aspect important de notre démarche est le fait de prendre en compte l'existence des familles multigéniques. Dans la deuxième partie nous proposons trois mesures, basées sur la distance de transposition dans le groupe symétrique, pour quantifier l'exceptionalité de l'ordre des gènes dans des régions génomiques conservées. Nous avons obtenu des expressions analytiques pour leur distribution dans le cas d'une permutation aléatoire. Dans la troisième partie nous avons étudié la distribution du nombre de cycles dans le graphe des points de rupture d'une permutation signée aléatoire. Nous avons utilisé la technique "Markov chain imbedding" pour obtenir cette distribution en terme d'un produit de matrices de transition d'une certaine chaîne de Markov finie. La connaissance de cette distribution fournit par la suite une très bonne approximation pour la distribution de la distance d'inversion.

Mots-Clés : Approximation de Poisson composée, méthode de Stein-Chen, région génomique conservée, test de significance, région de référence, familles multigéniques, comparaison de l'ordre des gènes, distance de transposition, permutation aléatoire, nombre de cycles dans le graphe des points de rupture, permutation signée aléatoire, Markov chain imbedding.

Abstract

This thesis is concentrated on some probability and statistical issues linked to genomic comparison. In the first part we present a compound Poisson approximation for computing probabilities involved in significance tests for conserved genomic regions found by the reference-region approach. An important aspect of our computations is the fact that we are taking into account the existence of multigene families. In the second part we propose three measures, based on the transposition distance in the symmetric group, for quantifying the exceptionality of the gene order in conserved genomic regions. We obtain analytic expressions for their distribution in the case of a random permutation. In the third part of the thesis we study the distribution of the number of cycles in the breakpoint graph of a random signed permutation. We use the Markov chain imbedding technique to obtain this distribution in terms of a product of transition matrices of a certain finite Markov chain. The knowledge of this distribution provides a very good approximation for the distribution of the reversal distance.

Keywords : Compound Poisson approximation, Stein-Chen method, conserved genomic region, significance test, reference-region approach, multigene families, gene order comparison, transposition distance, random permutation, number of cycles in the breakpoint graph, random signed permutation, Markov chain imbedding.

AMS 2000 subject classification : 60C05, 62E20, 62P10, 92D15.